AFRL-OSR-VA-TR-2015-0132

The role of benevolence in trust of autonomous systems

David Atkinson
FLORIDA INSTITUTE FOR HUMAN AND MACHINE COGNITION INC PENSACOLA FL

05/19/2015
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTC
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
15-04-2015	Final	15-4-2012 - 14-4-2015

4. TITLE AND SUBTITLE The role of benevolence in trust of autonomous systems	**5a. CONTRACT NUMBER** FA9550-12-1-0097
	5b. GRANT NUMBER FA9550-12-1-0097
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Atkinson, David J.	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) FLORIDA INSTITUTE FOR HUMAN AND MACHINE COGNITION INC 40 S ALCANIZ ST PENSACOLA FL 32502-6008	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) USAF, AFRL DUNS 143574726 AF OFFICE OF SCIENTIFIC RESEARCH 875 N. RANDOLPH ST. ROOM 3112 ARLINGTON VA 22203	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFOSR
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report provides a summary of the research and related activities performed. The issue of trust is one of the most significant obstacles to broad use of autonomy technology by DoD and other agencies. The impact of this research supports creation of computational methods that create a bridge to engineering of trustworthy autonomous systems. Objectives of this research were (1) to operationalize the quality of benevolence and understand how it contributes to well-calibrated trust of, and reliance upon, autonomous systems, (2) to investigate portrayal of trust-related attributes in the human-machine interface. Accomplishments include: (1) the formulation of benevolence as a complex belief structure with antecedent beliefs having semantic, temporal, causal and other interrelationships; (2) the mapping of a portion of this belief structure to measurable internal states of autonomous systems; (3) empirical evidence in support of the applicability of psychological concepts of interpersonal human trust to autonomous systems, including the role of personality and situation in modulating the role and importance of certain beliefs; (4) creation of a theory and engineering of a prototype Human Social Interface for machine portrayal of trust qualities in human-machine social interaction.

15. SUBJECT TERMS
Autonomy, Autonomous Systems, Robotics, Machine Intelligence, Trust, Human-Machine Interface, Human-Robot Interaction, Social Robotics, Trustworthy systems, Artificial Intelligence

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Dr. David J. Atkinson
U	U	U	UU	392	**19b. TELEPHONE NUMBER** *(Include area code)* 352-387-3063

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18
Adobe Professional 7.0

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

The Role of Benevolence in Trust of Autonomous Systems

FINAL REPORT

Contract Number: FA2386-11-1-4064

Air Force Office of Scientific Research

Prepared for

Benjamin Knott, PhD

Program Manager, Air Force Office of Scientific Research

Submitted by

David J. Atkinson, Ph.D.

Senior Research Scientist
Institute for Human and Machine Cognition
15 SE Osceola St. Ocala, FL 34471
datkinson@ihmc.us

15 April 2015

EXECUTIVE SUMMARY

This report is furnished to the Air Force Office of Scientific Research as the final deliverable for AFOSR Grant FA9550-12-1-0097, "The Role of Benevolence in Trust of Autonomous Systems." This report provides a summary of the research and related activities performed with the support of this grant and key results, including pre-print copies of the peer-reviewed publications and related material. Without reliable and robust methods for assessing the trustworthiness of intelligent, autonomous systems, the issue of trust has become one of the most significant obstacles to broad use of autonomy technology by DoD and other agencies. However, the impact of the research described in this report supports creation of computational methods that create a bridge to future engineering of trustworthy autonomous systems. The core objectives of this research were (1) to operationalize the quality of "benevolence" and understand how it contributes to well-calibrated trust of, and reliance upon, autonomous systems, and (2) to investigate portrayal of trust-related attributes in the human-machine interface. Significant headway was achieved on key topics, including some notable results. Key accomplishments discussed in this report include: (1) the formulation of benevolence as a complex "belief structure" with antecedent beliefs having important semantic, temporal, causal and other interrelationships; (2) the mapping of a portion of this belief structure to measurable internal states of autonomous systems, thereby potentially creating new opportunities for assessment of trustworthiness of such systems; (3) the obtaining of empirical evidence in support of the proposition that previous psychological concepts of interpersonal human trust are applicable to trust in autonomous systems, including the role of personality and situation in modulating the role and importance of certain beliefs; (4) the creation of a theory of a "Human Social Interface" which, when expressed in systems engineering terms, provides guidance for machine portrayal of trust-related qualities in human-machine social interaction; (5) the design and implementation of a software prototype based on the Human Social Interface theory that provides a basis for future experimentation and evaluation. This project resulted in eight peer-reviewed publications and sixteen presentations in scientific venues, meetings with distinguished visitors, and other in support of technology transition opportunities within DoD and to industry. Many new research questions were generated and there remains considerable work to do to fully understand the role of benevolence with respect to intelligent autonomous systems. Overall, the theoretical foundation for trustworthiness of autonomous systems is immature and remains an important area of focus for multiple disciplines.

TABLE OF CONTENTS

EXECUTIVE SUMMARY ii

TABLE OF CONTENTS iii

INTRODUCTION 1
NEEDS OF THE US AIR FORCE 1
OBJECTIVES 1
APPROACH 1
IMPACT 2

ACCOMPLISHMENTS 3

SUMMARY OF RESEARCH ACTIVITIES 5
THEORY DEVELOPMENT 5
EXPLORATORY SURVEY RESEARCH 5
PROTOCOL FOR STUDY OF ATTRIBUTION OF BENEVOLENCE 7

RESULTS AND FINDINGS 10
BENEVOLENCE AND TRUST 10
SURVEY RESEARCH ON TRUST-RELATED BELIEFS ABOUT AUTONOMOUS SYSTEMS 11
BELIEF STRUCTURES: ANTECEDENTS OF ATTRIBUTION OF BENEVOLENCE 14
THE HUMAN SOCIAL INTERFACE 19

LIST OF PUBLICATIONS 22

INVITED PRESENTATIONS 23

SIGNIFICANT EVENTS 24
NOTABLE TECHNOLOGY BRIEFINGS 24
ORGANIZATION AND PROGRAM ACTIVITIES 24
REVIEWER ACTIVITIES 25

APPENDICES 26

INTRODUCTION

This report is furnished to the Air Force Office of Scientific Research as the final deliverable for AFOSR Grant FA9550-12-1-0097, "The Role of Benevolence in Trust of Autonomous Systems." The research described here was performed between April 2012 and April 2015. The purpose of this report is to provide a summary of the research and related activities performed with the support of this grant and to summarize key results, including copies of the peer-reviewed publications and related documentary material. Please note that the publications found in the appendix here are author pre-print copies and may differ in some details from the published versions; the latter should be regarded as the scientific documentation of record.

NEEDS OF THE US AIR FORCE

The US Air Force forecasts the need to interact with and rely on increasingly intelligent autonomous systems. Without reliable and robust methods for assessing the trustworthiness of an autonomous system, the issue of trust has become one of the most significant obstacles to broad use of autonomy technology even as it rapidly matures.

OBJECTIVES

There were two primary objectives for the research, summarized as follows:

1. Operationalize "benevolence" and understand how that quality contributes to well-calibrated trust of, and reliance upon, autonomous systems.
2. Investigate measures and methods for portrayal of trust-related attributes such as "benevolence" in the human-machine interface.

APPROACH

The overall approach of this project was to relate empirically discovered trust-related qualities to theoretical constructs on which to base computational methods for further experimentation and development.

The research design was primarily qualitative and oriented towards establishing a theoretical framework for "benevolence" within which to examine the issues of trust and delegation to autonomous systems. In addition to theory development, this project included an exploratory survey to explore human attribution of trust-related qualities to autonomous systems and secondly, a human study to further refine the parameters of trust with a focus on the conditions under which an attribution of benevolence may occur. Analysis of the results, the publication and presentation of

study results in appropriate scientific venues, and this final report conclude the activities in this project.

IMPACT

The theory of trustworthiness in autonomous systems is immature. The scientific impact of operationalizing benevolence and component trust-related qualities using theoretical constructs from Cognitive Science and AI supports creation of computational methods to create a bridge to future engineering of trustworthy autonomous systems.

ACCOMPLISHMENTS

The principal scientific accomplishments of this project are summarized below and discussed in detail in the following sections.

- The attributed quality of "benevolence" to a candidate trustee (human or machine) was formulated as a construct consisting of a rich set of component beliefs with complex interrelations. These component beliefs, or "antecedents" of benevolence, have each been the focus of previous studies of human interpersonal trust. The formulation of benevolence arising from this study revealed the importance of perception of agency and animacy ("liveness") for autonomous systems.
- The project developed a semantic *belief structure* representation of trust qualities (component beliefs) for benevolence, including logical, temporal, causal, evidentiary relations and other dependencies among those beliefs and specified a preliminary mapping of those belief structure representations to facets of the internal state of autonomous systems. The objective of devising *new methods of measurement* of these internal states proved to be too difficult to complete given our current understanding and ability to analyze the internal state of autonomous systems. This is a topic for future research.
- The project obtained empirical evidence that confirmed certain key abstract qualities of human interpersonal trustworthiness (i.e., *Competence, Predictability, Openness, Risk/Safety*) are applicable to evaluation of autonomy trustworthiness. However, *self-reports* by study participants regarding the relative importance of trust-related qualities in the absence of specific context proved to be poor predictors of actual delegation decisions. The qualities most significantly related to evaluation of trustworthiness of an autonomous system, and their relative importance, varied by individual *personality* and *situational* factors (including, for example, *perception* and *acceptance of risks* of different types).
- The project formulated a theory of a *Human Social Interface* as an aid to engineering computational methods that portray anthropomorphic trust-related qualities in a human-machine cyber-physical interface. This formulation guided the design and programming of a software prototype for portrayal of trust-related qualities by a social robot in a second human study. This novel software architecture features a hybrid reactive/deliberative control scheme that enables loose coupling and non-interference of social

and task behaviors, and is easily extended as new social interactive requirements for autonomous robots are defined.

- The prototype Human Social Interface was tested in an immersive simulated environment designed to potentiate a heightened sense of danger. The simulation was designed to explore conditions under which attribution of benevolence might be important in a candidate autonomous robot application to urban rescue. A human study was designed, approved and implemented. However, trials for the study remained incomplete at the time of project expiration.

Accomplishments are discussed in more detail in the following sections. Peer-reviewed publications (attached) provide the documentation of record. Two additional papers and an invention report are in preparation.

SUMMARY OF RESEARCH ACTIVITIES

This section presents an overview of the key research activities performed under this grant. Results and other findings from these activities are described in the following section ("Summary of Findings").

THEORY DEVELOPMENT

Theory development in this project focused on development of a social model of trust that extends the cognitive and affective nature of human interpersonal trust in ways that will ultimately provide guidance for the design and development of autonomous agents that have the ability to engender appropriate human-machine reliance and interdependency. Furthermore, we defined and explored the concept of a "Human Social Interface" in system engineering terms rather than psychological terms. This theoretical aspect of the project was oriented towards the objective of devising computation mechanisms for modulation of human beliefs by a socially competent intelligent autonomous agent.

EXPLORATORY SURVEY RESEARCH

Previous research in psychology and other fields guided the specification of candidate Belief Structures for trust in autonomous systems. These in turn informed the design of our *survey research*, *robot simulation and human study protocol*, and *computational mechanisms*.

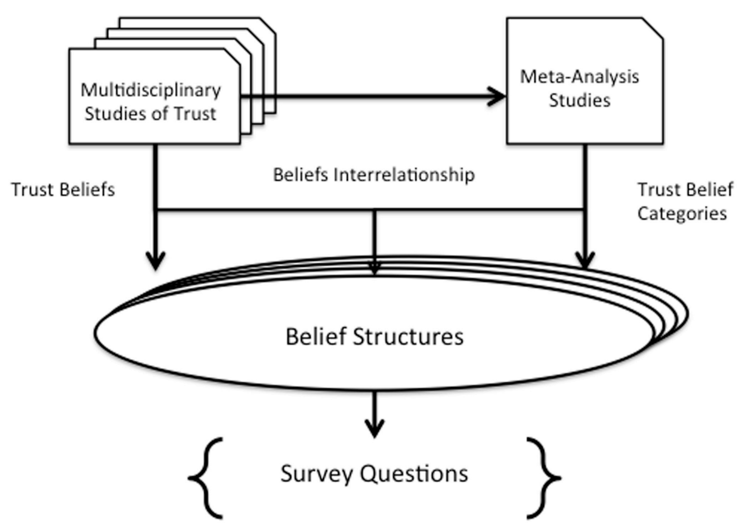


Figure 1: Flow of Information from Previous Studies

The trust-relevant internal state of a human agent is a complex cognitive and affective structure that references causal factors, attitudes, evaluations and

expectations centered on other agents (especially the potential "Trustee"), the situation, goals, and tasks. This state includes antecedent beliefs of the Trustor regarding a Trustee. Such beliefs may have complex interrelationships based on the relative contributions of evidence, causality and situational influences. To represent these beliefs and their interrelationships, we use the term "Belief Structure". The word "structure" explicitly reminds us that the individual beliefs in each belief structure are complex, inter-related, conditional, and occasionally may even be contradictory. Our ultimate goal of a computational representation of a belief structure must be rich enough to capture this logical structure.

An exploratory survey research study was performed to examine putative components of trust in autonomous agents. The study had three goals: 1) to assess which, if any, among a set of anthropomorphic beliefs derived from studies on human interpersonal trust are important to a human's decision to delegate to an intelligent, autonomous agent; 2) to determine the relative importance among such beliefs; 3) to explore whether the applicability or importance of those beliefs to a delegation decision vary in a systematic way by individual personality and/or situational factors.

The survey was conducted online with a sample of participants drawn from a pool of autonomy subject matter experts and decision-makers. This group was targeted specifically due to the essential role they play in developing autonomy technology and, more importantly, in helping to make decisions on whether to create and field specific types of applications. Thus, their attitudes and beliefs regarding trust of autonomous systems have the potential for broad effects on all aspects of the autonomous system lifecycle.

The survey design is summarized below. Please see the next section regarding findings and also refer to the publications located in the appendix for more information. The complete survey, as administered, was provided to Volkswagen Research of America, at their request, for use in research on the role of personality in trust of automobile autonomy.

Survey Design:

1. Participants ranked the absolute (not relative) importance of twenty-eight specific trust-related qualities of agents that span four Belief Structures defined by the theoretical portion of the project (Competence, Predictability, Openness, Safety). See Figure 2, below.

2. Participants completed three standard personality survey instruments used in the social sciences: Big Five Inventory (BFI-40), Individual Innovativeness (II), and the Domain- Specific Risk Taking Scale (DOSPERT).

3. Participants were presented with six challenge scenarios that vary in terms of risk type and magnitude as well as relative emphasis on aspects of the four focus Belief Structures. Participants were forced to make a choice of whether to rely on human, autonomous system, or "other" agent to satisfy the needs presented in the scenario. (Descriptions of these scenarios are included in the Appendix.)

4. Participants completed the Source Credibility inventory to explore the perceived "ethos" of the autonomous systems presented in the scenarios. This produced measures of perceived competence, goodwill, and overall trustworthiness.

Table 2 Twenty Eight Hypothetical Trust-Related Qualities of Intelligent, Autonomous Agents[b]

Category	Name	Quality Description
Competence	Capable	The autonomous agent can achieve a desired result.
	Knowledge	The autonomous agent has all the knowledge it needs to do its job.
	Accurate	What the autonomous agent believes to be true is actually true.
	Skilled	The autonomous agent possesses good methods for using its knowledge to do its task.
	Logical	The autonomous agent reasons correctly according to logic.
	Heuristic	When it cannot figure out something using logic, the autonomous agent can make good guesses.
	Corrective	The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know.
	Adaptive	The autonomous agent learns to correct its mistakes, as well as to improve and maximize its capability.
Predictability	Expected	The autonomous agent's behavior conforms to expectations.
	Purposeful	The autonomous agent purposefully acts to achieve goals.
	Helpful	The autonomous agent will assist people, whenever it is possible.
	Directable	The autonomous agent accepts and carries out orders.
	Reasonable	The autonomous agent uses its knowledge and skills in expected ways.
Safety	Safe	The autonomous agent's behavior will not harm humans or human interests.
	Limited	Any incorrect behavior by the autonomous agent will not cause harm.
	Stable	The autonomous agent fails gracefully and recovers from its failure promptly.
	Ruled	The autonomous agent adheres to obligations, principles, and rules.
	Correctable	The autonomous agent can correct its own defects or they can be corrected by a human.
	Protective	The autonomous agent recognizes and avoids harming humans' interests.
	Favorable	Given alternatives in what to do or how to do it, an autonomous agent will act in a way that is favorable to a human being who might be affected.
Openness	Visible	What the autonomous agent is doing and how it works is easy to see and understand.
	Honest	The autonomous agent believes what it says.
	Transparent	It is easy to inspect an autonomous agent.
	Communicative	The autonomous agent communicates in a way that is easy to understand.
	Interactive	The autonomous agent responds when you are trying to communicate with it.
	Attentive	The autonomous agent is aware of communication between others nearby.
	Reactive	The autonomous agent responds quickly to calls for attention.
	Disclosing	The autonomous agent communicates truthfully and fully.

[b] Quality Description is shown exactly as it appeared in the survey.

Figure 2: Hypothetical Trust-Related Qualities

PROTOCOL FOR STUDY OF ATTRIBUTION OF BENEVOLENCE

The project developed an experimental protocol under which the attribution of benevolence by a person to an autonomous robot could be examined. As part of the apparatus for this study, a detailed immersive 3D simulation was constructed, including situation and environment conditions as well as emulated autonomous robots. The simulated situation and environment were designed to evoke a sense of danger in participants, potentiate "victim psychology," and thereby establishing the conditions in which benevolence can become most important (See Figures 3 and 4). From the perspective of potential applications, the key question was whether the

perceived benevolence of an intelligent robot would increase cooperation and compliance with an offer of help from the robot (e.g., rescue).

The emulation of robots within the simulation employed a prototype computational version of the Human Social Interface to control robot social behaviors during the study. In addition, software was developed for control of the trial. This included task sequence of events, a special interface for collection of pre- and post-task questionnaire data, and data acquisition within the simulated world. This software is Open Source and will be made available to other researchers.

The study design and methodology was first presented at ACM/IEEE HAI-14 in Japan:

- Atkinson, D.J. and Clark, M.H. Methodology for Study of Human-Robot Social Interaction in Dangerous Situations. In *Proceedings of Human-Agent Interaction.* DOI: 10.1145/2658861.2658871. ACM (2014).

(See figures, next page)

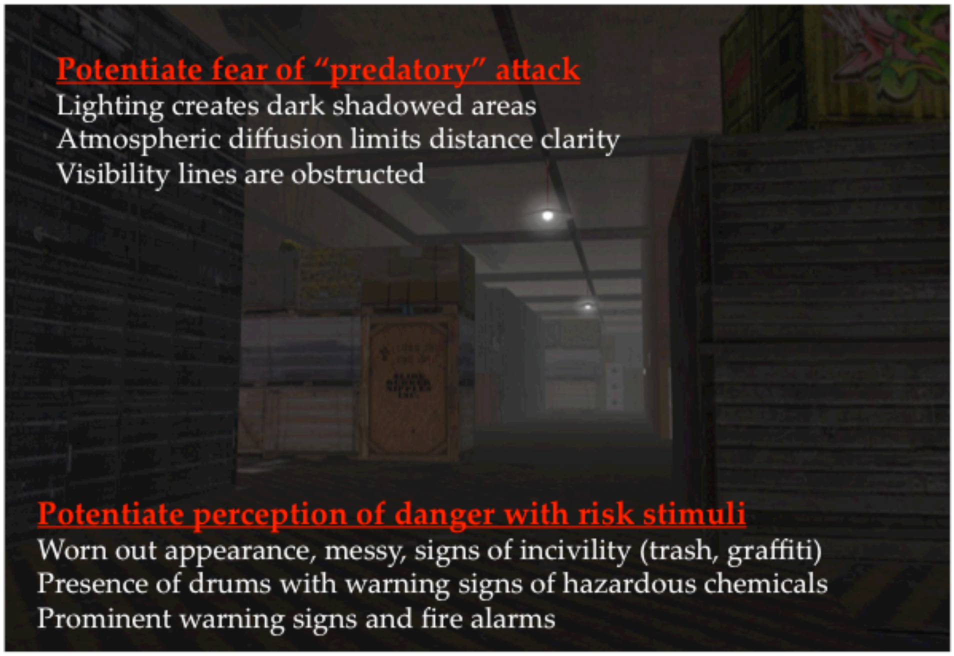


Figure 3: Simulated Environment Cues to Evoke Danger

Figure 4: Participant POV in Simulated Warehouse Disaster

RESULTS AND FINDINGS

This section presents an overview of results and other findings produced by the activities described in the previous section. It should be viewed as a guide to the full, detailed results as formally documented in peer-reviewed publications and related presentations. These are included in the Appendix to this report. Two additional publications are currently in preparation.

BENEVOLENCE AND TRUST

Trust is not benevolence and benevolence is not trust. The two concepts have been frequently conflated in the history of social psychology. Attribution of trustworthiness and benevolence are both products of entangled cognitive and emotional processes (Cummings and Bromily 1996). A belief that a potential Trustee is benevolent entails many of the same beliefs that are also *antecedents to trust*. However, some of the unique characteristics of benevolence, as opposed to trustworthiness in general, suggest that we must be careful to consider the constructs individually.

Believing that a Trustee is benevolent requires antecedent beliefs about the Trustee, such as those regarding motivation, dispositions and intentions, predictability, competence, and more. Furthermore, these antecedents may have complex interrelationships based on the relative contributions of evidence, causality and situational influences. These are discussed in more detail below in the section called "Belief Structures: Antecedent Beliefs for Attribution of Benevolence".

Importance of Benevolence to Applications of Autonomy

Benevolence appears to be an important quality for certain roles that we would like to target for application of autonomy technology. For example, there is a demand for applications of intelligent robotics in domains and situations that may be dangerous for humans, i.e., where there exist manifestly real or perceived threats to life or limb. Such threats may be due to *environmental* factors one might find in broad-area natural disasters (e.g., earthquakes) and local crises (e.g., urban structure fires). Threats may also be a result of *adversarial* factors due to armed conflict or crime. A prominent application of intelligent robots in dangerous situations that crosses both DoD and Civil interests is search and rescue.

Benevolence in the context of danger is of special interest for human-robot interaction because perceived danger evokes unique human psycho-physiological factors that influence perception, cognition and behavior (including interaction). This is referred to as "victim psychology". Human first responders who provide aid

to victims must contend with the abnormal psychology that such high-risk situations evoke; indeed, they receive special training for exactly this purpose. In the case of rescue, it is sometimes the case that a victim will not cooperate with the rescuer unless a great deal of trust has been established. One component that distinguishes such cases is that the rescuer is perceived to be putting himself at risk for the express benefit of helping a victim. As discussed below, this is a core attribute of benevolence. The design of intelligent robots for application in such situations must account for these psychological factors.

These results are discussed in the paper:

- Atkinson, D.J. and Clark, M.H. Methodology for Study of Human-Robot Social Interaction in Dangerous Situations. In *Proceedings of Human-Agent Interaction.* DOI: 10.1145/2658861.2658871. ACM (2014).

SURVEY RESEARCH ON TRUST-RELATED BELIEFS ABOUT AUTONOMOUS SYSTEMS

These results are discussed in the paper:

- Atkinson, D.J. and Clark. Anthropomorphism and Trust of Intelligent, Autonomous Agents by Early Adopters. Revision under Editorial Review for *International Journal of Social Robotics (SORO).* Springer (expected 2015).

Result (1)

The agent qualities **self-reported** as the most important for delegation (Figure 5) are *consistent with previous results* from interpersonal trust studies, i.e., (1) the ability of the machine to achieve the desired results, and (2) not causing harm.

Table 3 Top Three Most Important Autonomous Agent Qualities Reported by Participants

Rank	Name	Quality Description
1st	Safe	The autonomous agent's behavior will not harm humans or human interests.
2nd	Capable	The autonomous agent can achieve a desired result.
3rd	Limited	Any incorrect behavior by the autonomous agent will not cause harm.

Figure 5: Top Three Reported Important Qualities for Autonomous Agents

Result (2)

The self-reported important qualities of an autonomous system trustee for delegation were *not significantly correlated* with the actual choices for delegation

when participants considered specific use-case scenarios as shown in Figure 6. The names of the scenarios are shown as column headings. Several of these correlations are statistically significant at 98% confidence or better.

Table 4 Importance of Qualities of Autonomous Agent Significantly Correlated with Actual Participant Reliance on Autonomous Agent[c]

Airport Trans.	Financial Man.	Medical Proc.	Home Health.	Disaster Resp.	Lost at Sea
Corrective, $r = 0.396$	Accurate, $r = -0.405$	*none*	Visible, $r = 0.437$*	Corrective, $r = 0.418$* Heuristic, $r = 0.395$ Attentive, $r = 0.393$	Protective, $r = 0.419$* Visible, $r = -0.390$ Disclosing, $r = 0.375$

[c] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df $= 29$; * indicates $\alpha < 0.02$.

Figure 6: Actual Trust-Related Qualities Correlated with Delegation

When considered across scenarios, the specific qualities of autonomous systems, and the parent Belief Structures related to those qualities, are raised or lowered in importance depending on both situational (use-case specific) factors and individual psychological differences (Figure 7).

Table 6 Correlation of Perceived Risk and Benefit with Choice of Autonomous Agent by Scenario[e]

Scenario	Risk	Benefit
Airport Trans.	$r = -0.546$**	NS
Financial Man.	NS	NS
Medical Proc.	$r = -0.380$	$r = 0.585$**
Home Health.	$r = -0.470$**	$r = 0.632$**
Disaster Resp.	$r = -0.387$	$r = 0.484$**
Lost at Sea	NS	$r = 0.555$**

[e] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df $= 29$.
** indicates $\alpha < 0.01$.

Figure 7: Correlation of Perceived Risk and Benefit with Delegation Decision

In particular, individual high scores for *Extraversion*, *Openness*, and *Conscientiousness* were shown to be very important in some situations while in others, the *tolerance for risk* of certain types is dominant in decisions to rely upon an intelligent, autonomous agent (Figure 8).

Table 8 Participant Personality Factors Significantly Correlated with Reliance on Autonomous Agent[g]

Scenario	Correlated Personality Factor(s)	
Airport Trans.		*none*
Financial Man.	Innovation II, $r =$	0.355
Medical Proc.	BFI *Extraversion*, $r =$	0.368
	BFI *Openness*, $r =$	0.366
Home Health.	DOSPERT *Social Risk*, $r =$	0.364
Disaster Resp.	BFI *Conscientiousness*, $r =$	0.366
Lost at Sea	Innovation II, $r =$	-0.366

[g] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df $= 29$.

Figure 8: Personality Factors Correlated with Delegation

Summary of Key Survey Findings (95% Confidence)

1. We did not confirm any scenario-independent specific agent qualities that uniformly contributed to an affirmative human reliance decision. Certain qualities were important in some scenarios and not in others. Individual intuitions about "important" autonomous system trust-related attributes were uncorrelated with actual reliance choices in specific application scenarios. Our interpretation is that there may exist an influential disposition of beliefs regarding trustworthiness that are not necessarily the most salient during conscious introspection but become revealed when people are forced to apply their trust-related beliefs in challenging delegation decisions.

2. Specific qualities of agents, and categories of those qualities, are likely to be raised or lowered in importance depending on both situational (application-specific) factors and human psychological factors. Further investigation is required to identify those qualities and provide a mechanism for understanding how their role and importance in human reliance decisions changes.

3. Certain personality factors, including high scores for *Extraversion, Openness,* and *Conscientiousness* are very important in some situations while in others, the tolerance for certain kinds of risk is dominant when it comes to deciding to become reliant on an intelligent, autonomous system. This suggests that certain people may more readily accept a dependency on an autonomous system, and conversely, others are likely to be extremely resistant. Heretofore, most research on human-robot interaction has not considered the importance of individual differences.

BELIEF STRUCTURES: ANTECEDENTS OF ATTRIBUTION OF BENEVOLENCE

This section provides background and an overview of material that will be published in a future paper that is currently in development. The objective of this section is to provide insight specifically into the composition and representation of the components of "benevolence" in our operationalized formulation of the construct.

An attribution of benevolence requires the Trustor to hold certain beliefs about the Trustee and beliefs about the potential influence of external factors. Even the perception of superficial qualities that are at best heuristic indicators of trustworthiness can influence the attribution of benevolence, such as similarity of the Trustee's attitudes and preferences to those of the Trustor (Berschid and Walster; Newcomb).

The quality of similarity is by itself not sufficient for attribution of benevolence. Rather, it serves as an amplifier. A second amplifier for attribution of benevolence is observation or inference that the Trustee is acting in a way that is antithetical to the Trustee's self-interest (Holmes 1981).

However, certain beliefs about the Trustee are necessarily required and sufficient for attribution of benevolence. These span conceptual types that include affect, expectancy, and intentionality as determined by inference from the observed behavior of the Trustee or evidence from other sources (McKnight and Chervany 2001). Without any of these, the threshold for benevolence is not met. In that event, it devolves into another attributed quality, e.g., helpfulness.

Specific antecedent beliefs about the Trustee may have complex interrelationships based on the relative contributions of evidence, causality and situational influences. To represent these beliefs and their interrelationships, we use the term "Belief Structure". The word "structure" explicitly reminds us that the individual beliefs in each belief structure are complex, inter-related, conditional, and occasionally may even be contradictory. Our ultimate goal of a computational representation of a belief structure must be rich enough to capture this logical structure.

For a more detailed discussion of Belief Structures, please see the publication mentioned earlier. A review copy is attached as an appendix to this report:

- Atkinson, D.J. and Clark. Anthropomorphism and Trust of Intelligent, Autonomous Agents by Early Adopters. Revision under Editorial Review for *International Journal of Social Robotics (SORO).* Springer (expected 2015).

The following paragraphs provide an overview of the component antecedents of the "Benevolence" Belief Structure. Concept Maps of each component are shown.

Goodwill

Goodwill has been examined in detail by numerous studies. We've adopted the general form described by McGrosky and Tevin (1999).

Disposition to sympathy or concern with needs of other

This disposition entails the perception of a positive orientation of the Trustee towards the Trustor (Mayer, Davis and Schoorman 1995). That is, the Trustee wishes the Trustor well. Benevolence in trust involves the perceived willingness of the Trustee to behave in a way that benefits the interests of both parties with a genuine concern for the partner despite certain risks or loss (Hui 2005).

Absence of opportunism

The absence of opportunism entails a belief that the Trustee has nothing to gain, that is, there is no desire to capitalize on the dependency created by Trustor as a consequence of trust.

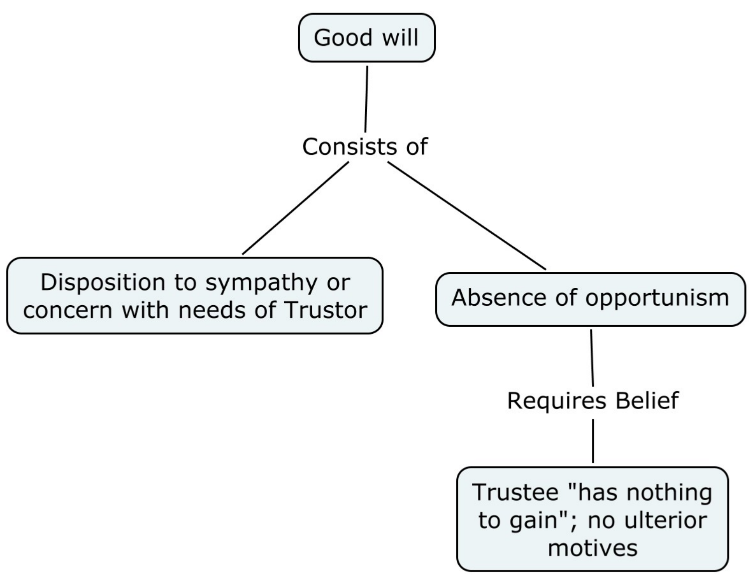


Figure 9: Concept Map for "Goodwill"

No hidden ill will

Sometimes this is referred to as the quality of "integrity", however that term is ambiguous without considerable discussion.

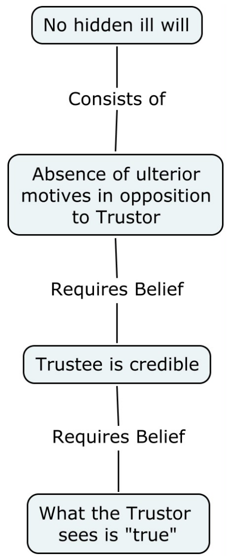


Figure 10: Concept Map for "No Hidden Ill Will"

The Trustee is expected to behave in a manner that benefits the "dependent" Trustor even when doing so is antithetical to the Trustee's self interest (Kelly 2003). This belief about the character of the Trustee is most important in early evaluation by the Trustee, and while this period may be quite short it is a strong filter that can derail trust quickly; a judgment of "ill will" corresponds to a perception of high risk. This belief's contribution to attribution of benevolence decreases relative to other component beliefs as more evidence about the Trustee is obtained.

Absence of ulterior motives

The aspect of specific relevance to benevolence is the belief that a Trustee harbors no hidden desire or intent to harm the interests of the Trustor. This is the same or a very similar component belief of "Goodwill" called "Absence of Opportunism". The more confidently observers can infer that Trustee behavior is antithetical to self-interest, the stronger the attributions of benevolence (Holmes 1981).

Trustee is credible

The credibility of the Trustee is essential, that is, what the Trustor sees is "true". If the Trustee is perceived to have a motivation to lie, credibility is greatly reduced (Hovland, Janis and Kelly 1953). The salience and importance of this belief is particularly important early in a trust relationship when interaction with the Trustee helps the Trustor to gain insight (Mayer, Davis and Schoorman 1995).

Intentional Social Relationship

Benevolence depends upon the entering of two actors into a pro-social relationship wherein there is an expectation of a bilateral disposition and intention to act in a

way that achieves some degree of mutual goal adoption and achievement. It is a bilateral dependency but it is not necessarily symmetrical. Attribution of benevolence by one actor towards the other requires particular beliefs about the intentionality of the other actor, including the persistence of those beliefs under the prevailing conditions.

The "Disposition to Act Favorably" (See Figure 11) can be thought of as an arrangement of factors that potentiate action. In motivational terms, it is an inclination or tendency. In cognitive terms, it represents an increased likelihood of something (belief, goal, intention, action). The "Disposition to Act Favorably" *presupposes a mechanism for choice* (agency). It also requires that choice is not random but can be biased towards certain outcomes.

The "Intention to Act Favorably" is an active state of goal pursuit. It is a belief that, given the "right" circumstances, the trustee will act (See Figure 12).

The "Predictability, Persistence and Stability of Intentions" refers to the belief that Trustee actions are consistent enough to be forecasted in a given situation. Although circumstances might prevent the Trustee from favorable action, the Trustee is inclined not to change the disposition or intention to act favorably.

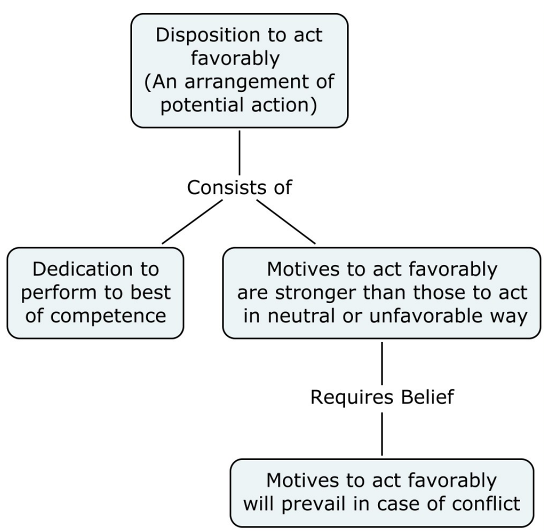


Figure 11: Concept Map for "Disposition to Act Favorably"

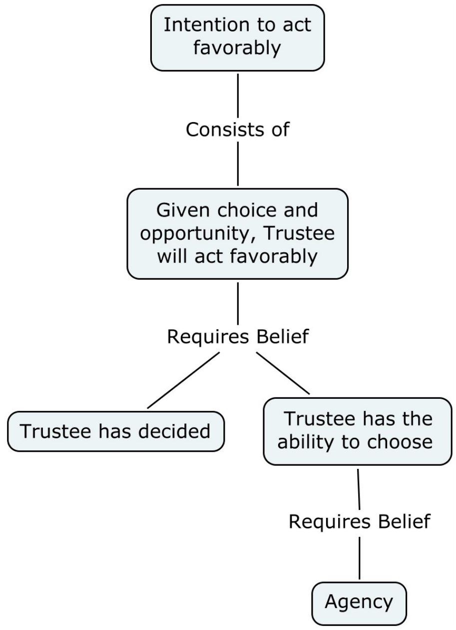


Figure 12: Concept Map for "Intention to Act Favorably"

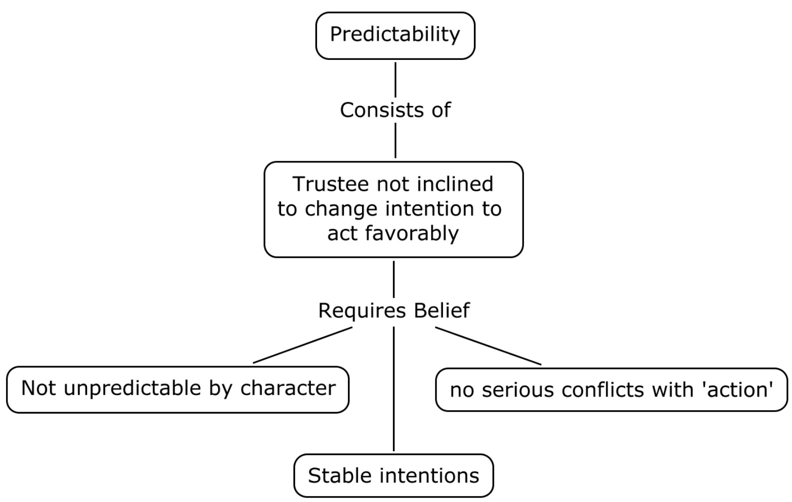


Figure 13: Concept Map for "Predictability"

Competence

The Competence Belief Structure reflects the Trustee's expertise, ability, skills, aptitude and so forth. It also requires a consideration of context, i.e., that the Trustee can apply his competence to achieve a result of value to the Trustor.

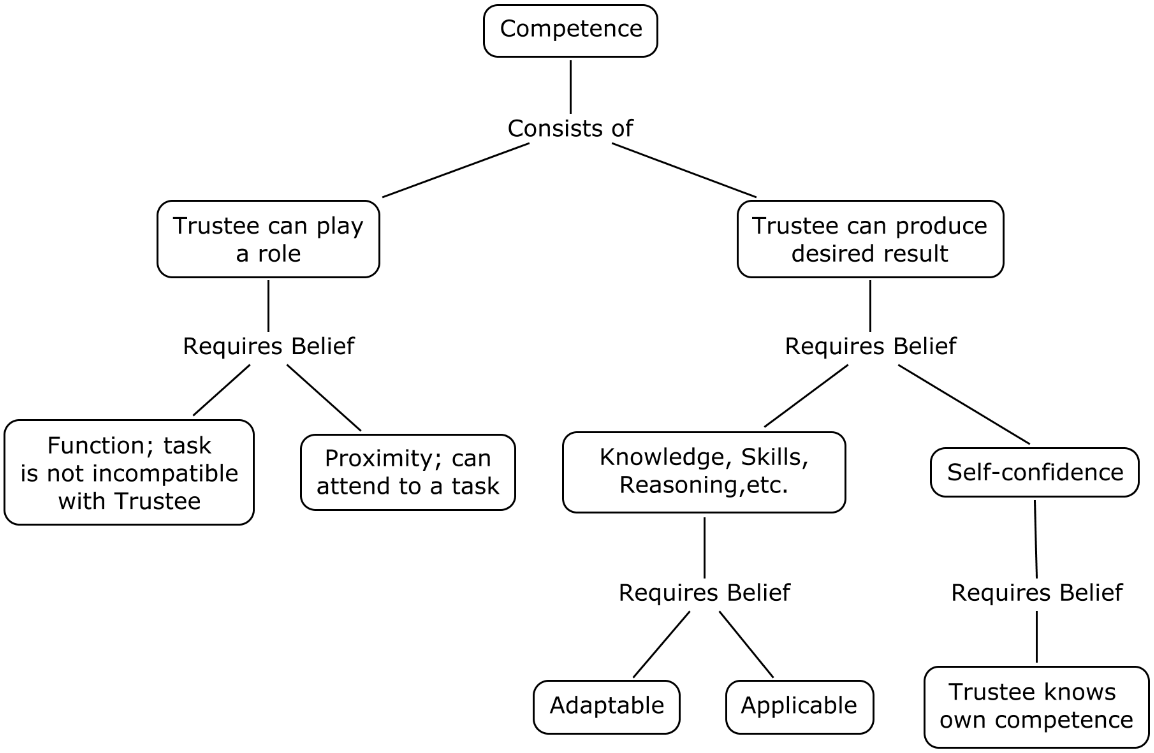


Figure 14: Concept Map for "Competence"

THE HUMAN SOCIAL INTERFACE

The project formulated a theory of a *Human Social Interface* for engineering computational methods to portray anthropomorphic trust-related qualities in human-machine cyber-physical interfaces. This formulation is prescriptive, not descriptive, and thereby provides a practical systems engineering guide for future system developers of autonomous systems that is based on an established psychological foundation. From an engineering perspective, the purpose of trust in a multi-agent system composed of human and machine elements is to achieve optimal overall performance via appropriate interdependency, mutual reliance, and appropriate exchange of initiative and control between the cognitive components (human and/or machine). See the following publication in the Appendix for more details:

- Atkinson, D. J., and Clark, M. H. Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust. *In Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium.* Technical Report No. SS-13-07. Menlo Park: AAAI Press (2013).

A specification of this interface between human and machine describes:

- *Assumptions* about each agent (as part of a Theory of Mind)
- *Communicative signals* (e.g., relative position and orientation) on specific information channels (e.g., proxemics)
- *Interaction protocols* that specify how and when signals and channels are used in specific (operational) contexts, and how the internal state of each agent consequentially changes in response to such signals.

See the illustration shown in Figure 15, below. A detailed discussion and presentation of how the Human Social Interface is used in social robotics is the subject of the second paper currently in preparation.

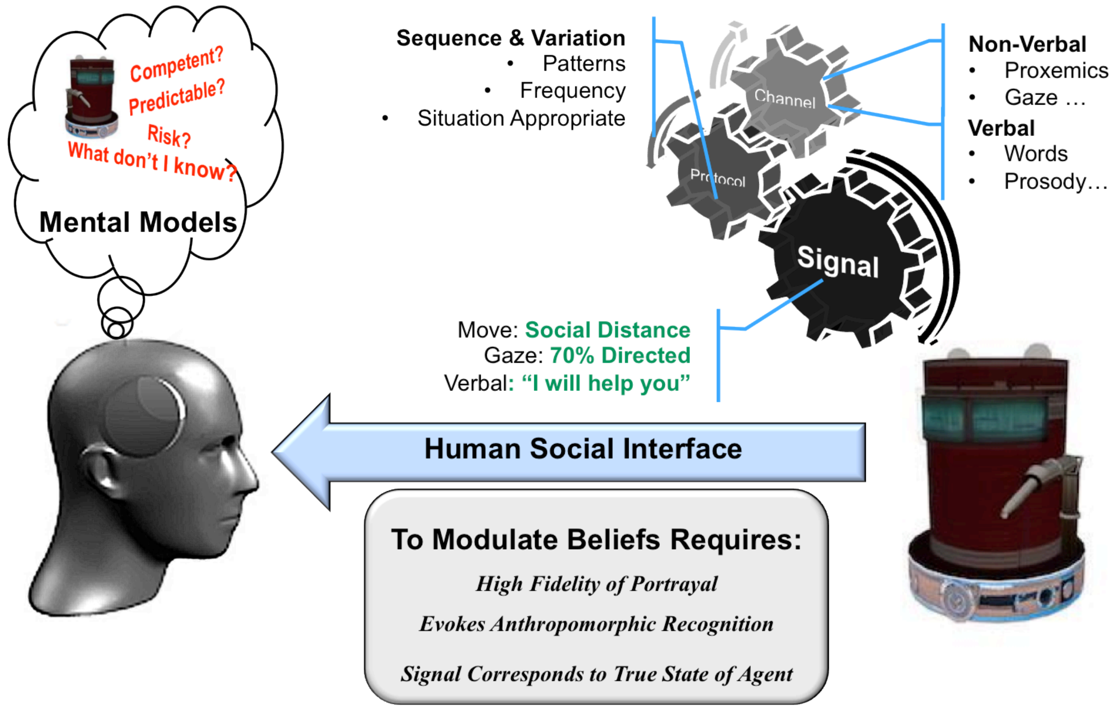


Figure 15: Elements of the Human Social Interface

REFERENCES

Citations found in the narrative above are listed here. Please see each of the publications produced by this project for a complete list of relevant sources to this project as a whole.

Berscheid, E., & Walster, E. H. (1978). *Interpersonal Attraction.* 2nd edition. Reading, MA: Addison-Wesley.

Castelfranchi, C. (2008) "Trust and reciprocity: misunderstandings," *International Review of Economics*, vol. 55, no. 1, pp. 45–63.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots.

Robotics and Autonomous Systems, 42, 143-166. doi: 10.1016/S0921-8890(02)00372-X

Holmes, J. G. (1991) Trust and the Appraisal Process in Close Relationships. In Jones, W. H. , Perlman, D. (eds.): *Advances in Personal Relationships*. Jessica Kingsley, London 2:57-104

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and Persuasion*. New Haven, CT: Yale University Press.

Hui, M. K. and Sit, A. Y. (2005) "Service failure and trust: the roles of benevolence, competence, and culture," presented at the Services Systems and Services Management. Proceedings of ICSSSM '05. 2005 International Conference on, 2005, vol. 1, pp. 180–182.

Kelly, Harold H. (2003) *An Atlas of Interpersonal Situations*. Cambridge University Press. Pp. 258-259.

Larzelere, R., & Huston, T. (1980). The dyadic trust scale: Toward understanding interpersonal trust in close relationships. *Journal of Marriage and the Family*. 42:595-604.

Mayer, Roger C., Davis, James H. and Schoorman, F. David (1995) An Integrative Model of Organizational Trust. *The Academy of Management Review*. 20(3):709-734.

McCroskey, J. C. and Teven, J. J. (1999)"Goodwill: A reexamination of the construct and its measurement," *Communications Monographs*.

McKnight, D. H. and Chervany, N. L. (2001) "What Trust Means in E-Commerce Customer Relationships: An Interdisciplinary Conceptual Typology," *International Journal of Electronic Commerce*, vol. 6, no. 2.

Minato, T., Shimada, M., Ishiguro, H., & Itakura, S. (2004). Development of an android robot for studying human-robot interaction. *Lecture Notes In Computer Science: Proceedings of the 17th international conference on Innovations in applied artificial intelligence, 17*(20), 424-434.

Newcomb, T. M. (1956). The prediction of interpersonal attraction. *American Psychologist* 11:575-586.

Nooteboom, B. (2002), *Trust: Forms, foundations, functions, failures and figures*. Cheltenham: Edward Elgar.

Rosen, B., & Jerdee, T. H. (1977). Influence of subordinate characteristics on trust and use of participative decision strategies in a management simulation. *Journal of Applied Psychology*, 62:628-631.

LIST OF PUBLICATIONS

The project provided essential support in whole or part for preparation of peer-reviewed publications, presentation of the results in scientific venues, and related professional activities of the Principal Investigator. The list of publications is presented below in reverse chronological order. See the appendix for author pre-print copies of the full papers. Two additional papers are in preparation but not yet submitted for publication.

- Atkinson, D.J. and Clark. Anthropomorphism and Trust of Intelligent, Autonomous Agents by Early Adopters. Final revision under Editorial Review for *International Journal of Social Robotics (SORO).* Springer (expected 2015).
- Atkinson, D.J. Emerging Cyber-Security Issues of Autonomy and the Psychopathology of Intelligent Machines. In *Foundations of Autonomy, Papers from the 2015 AAAI Spring Symposium on*. AAAI. Menlo Park: AAAI Press (2015).
- Atkinson, D.J., Dorr, B.J., Clark, M.H., Clancey, W.J., Wilks, Y. Ambient Personal Environment Experiment (APEX): A Cyber-Human Prosthetic for Mental, Physical and Age-Related Disabilities. In *Ambient Intelligence for Health and Cognitive Enhancement, Papers from the 2015 AAAI Spring Symposium on.* AAAI. Menlo Park: AAAI Press (2015).
- Atkinson, D.J. Robot Trustworthiness: Guidelines for Simulated Emotion. In *HRI '15: ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts Proceedings.* ACM (2015).
- Atkinson, D.J., Clancey, W.J. and Clark, M. Shared Awareness, Autonomy and Trust in Human-Robot Teamwork. *In Artificial Intelligence for Human-Robot Interaction. Papers from the 2014 AAAI Fall Symposium. Technical Report No. FS-14-01.* Menlo Park: AAAI Press (2014).
- Atkinson, D.J. and Clark, M.H. Methodology for Study of Human-Robot Social Interaction in Dangerous Situations. In *Proceedings of Human-Agent Interaction.* DOI: 10.1145/2658861.2658871. ACM (2014).
- Atkinson, D. J., and Clark, M. H. Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust. *In Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium.* Technical Report No. SS-13-07. Menlo Park: AAAI Press (2013).
- Atkinson, David J., Friedland, Peter and Lyons, Joseph B. Human-Machine Trust for Robust Autonomous Systems. In *Proceedings of IEEE Human-Robot Interaction Conference (HRI-12). IEEE Workshop on Human-Agent-Robot Teamwork*. IEEE Press (2012)

INVITED PRESENTATIONS

The preparation and presentation of a number of invited presentations was supported in part by this grant. The list is presented below in reverse chronological order. See the Appendix for titles, charts and posters used in these presentations.

• Poster (Late Breaking Report), Conf. Human Robot Interaction, ACM/IEEE	Mar 2015
• Participant, Symposium on Ambient Intelligence for Health and Cognitive Enhancement, AAAI Spring Symposium Series	Mar 2015
• Participant, Symposium on Foundations of Autonomy and Its (Cyber) Threats, AAAI Spring Symposium Series	Mar 2015
• Participant, Symposium Artificial Intelligence for Human-Robot Interaction, AAAI Fall Symposium Series	Nov 2014
• Participant, ACM/IEEE Human-Agent Interaction	Aug 2014
• Lecturer, Robotics Camp (Science outreach to secondary school students)	Jul 2014
• Lecturer, Computer Science Department, Tulane University	Feb 2014
• Keynote, IEEE/ACM International Conference on Intelligent Agent Technology and Web Intelligence	Nov 2013
• Participant, Symposium on Trust in Autonomous Systems, AAAI Spring Symposium Series	Mar 2013

SIGNIFICANT EVENTS

In addition to invited scientific presentations, this project supported the participation of the PI in a number of related professional activities related to the research addressed by the project. Several of these can be classed as "technology transition" activities of programmatic importance to AFRL/AFOSR. The significant events included professional and program-related activities, peer-review, and individual discussions and briefings by the PI to notable visitors and other government officials.

NOTABLE TECHNOLOGY BRIEFINGS

• Prof. Larry Leifer and Staff, Stanford Univ., Volkswagen Automotive Interface Laboratory ***Technology Transition event**	Mar 2015
• Research Staff at SoarTech, Inc. in support of their SBIR proposal to AFRL/RX ***Technology Transition event**	Mar 2014
• Dr. Melissa Flagg, Office of Secretary of Defense, DSO	Feb 2013
• Mr. James Overholt, Chief Scientist for Autonomy, AFRL/RI-711HPW	Nov 2013
• Prof. Soo-Young Lee, Director Brain Science, KAIST, Korea	Sep 2013
• Dr. Robert Neches, Director IA/IARPA	Jun 2013
• Mr. Adam Lovelace, Office of Secretary of Defense, DTI	Oct 2012

ORGANIZATION AND PROGRAM ACTIVITIES

• Program Committee, Foundations of Autonomy and Its (Cyber) Threats, 2015 AAAI Spring Symposium	2015
• The IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)	2014 - 2015
• Program Committee, IEEE/WIC International Conference on Intelligent Agent Technology	2014 - 2015
• General Chair, AFOSR Workshop on Human-Machine Trust for Robust Autonomous Systems, Ocala, FL	2012
• Session Chair, Research Challenges, NASA Workshop on Autonomy Validation, Pasadena, CA ***Technology Transition event**	2012
• Session Chair, Research Challenges, AFRL Workshop on Human Centered Autonomy Dayton, OH * **Technology Transition event**	2012

REVIEWER ACTIVITIES

• AI Magazine	2015
• Young Pioneers, 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2015)	2015
• The Intersection of Robust Intelligence and Trust in Autonomous Systems, 2014 AAAI Spring Symposium	2014
• Artificial Intelligence for Human-Robot Interaction, 2014 AAAI Fall Symposium	2014
• IEEE Transactions on Robotics	2014
• Air Force Office of Scientific Research	2012 - 2014
• 7th FINT Workshop on Trust Within and Between Organizations	2013

APPENDICES

APPENDIX A. STUDY PROTOCOLS AS APPROVED

APPENDIX B. AUTHOR COPY OF EACH PUBLICATION

APPENDIX C. SELECTED PRESENTATIONS

APPENDIX D. PROGRAM REVIEW CHARTS

APPENDIX E. ADDITIONAL TECHNICAL MATERIAL

APPENDIX A. STUDY PROTOCOLS AS APPROVED

- **Survey on Autonomous Agents**
- **The Role of Benevolence in Trust of Autonomous Systems**

IHMC IRB SCREENING FORM
Committee on Human Participation in Research

Researcher: GO TO PAGE 3. Page 1 is NOT to be completed by the Researcher.

DATE OF RECEIPT OF PROPOSAL:	3 August 2012
PRINCIPAL INVESTIGATOR:	David Atkinson
TITLE OF PROJECT:	Survey on Autonomous Agents
SPECIAL CONSIDERATIONS (i.e., generic proposal, expedited review, other special considerations)	Possible waiver of requirement for written informed consent. Qualifies for expedited review. RRHoffman 13Aug2012

REVIEW BY INDIVIDUAL IRB MEMBER #1

REVIEWER NAME	JOHN W. COFFEY
SIGNATURE	John W. Coffey
DATE	8/14/2012
DECISION (Approve, Disapprove, Approve pending modifications)	Approve
COMMENTS	

REVIEW BY INDIVIDUAL IRB MEMBER #2

REVIEWER NAME	
SIGNATURE	
DATE	
DECISION (Approve, Disapprove, Approve pending modifications)	
COMMENTS	

REVIEW BY INDIVIDUAL IRB MEMBER #3

REVIEWER NAME	
SIGNATURE	
DATE	
DECISION (Approve, Disapprove, Approve pending modifications)	
COMMENTS	

REVIEW BY INDIVIDUAL IRB MEMBER #4

REVIEWER NAME	
SIGNATURE	
DATE	
DECISION (Approve, Disapprove, Approve pending modifications)	
COMMENTS	

REVIEW BY MEMBER # 5 - IRB CHAIR

IRB Chairperson's Signature	
DATE	
IRB DECISION (Approve, Disapprove, Approve pending modifications)	
COMMENTS	

SIGNATORY OFFICIAL

Signatory Official's Signature	

DATE	August 15, 2012
FINAL DETERMINATION	
COMMENTS	

PI NAME: David J. Atkinson
STUDY TITLE Survey on Autonomous Agents GRANT TITLE: The Role of Benevolence in Trust of Autonomous Systems
BRIEF DESCRIPTION OF PROJECT'S PURPOSES This submittal is for the first of several studies under this grant. The purpose of this study, an online survey, is exploratory: to elicit attitudes regarding the relative importance of a set of traits/characteristics/properties of autonomous systems to trust-related decisions, and; to identify any potentially interesting correlations between these attitudes, personality, and/or demographic grouping(s). Unlike previous studies that have focused on interpersonal trust, or human trust of automation, this study will focus exclusively on machines with a high degree of intelligence and autonomy to choose courses of action to accomplish complex objectives. The results of this study will inform the development of designs for laboratory experiments to be conducted in the next year of this project. Those laboratory experiments will be submitted for IRB approval when the design is appropriately mature.
PLANNED DATES FOR INITIATION AND COMPLETION OF THE PROJECT Data gathering via the survey is planned for 1 September 2012 through 31 October 2012
NUMBER and CHARACTERISTICS OF PARTICIPANTS (e.g., target population, age range, anticipated male/female ratio, ratio of minorities, special populations, etc.) Target population consists of adults between the ages of 21 and 60 with a total number sufficient to control for individual differences as required. We are hoping for about 100 valid responses. Participants will have a mix of educational and career backgrounds. We are specifically targeting individuals who have an interest in autonomous systems, including educational, prospective users, developers, decision-makers, and others who could be affected by future applications of this technology. However, it is possible that participants may or may not have any knowledge or interest in the subject.

following expression of interest in the survey by the individual [clarification].

METHOD OF RECRUITMENT (note any tangible or intangible benefits, monetary payments, or other inducements)

The survey will include individuals in several groups, recruited by:

1. Personal telephone and or email contact by the PI. These are individuals identified as potentially having some degree of potential interest in autonomous systems on the basis of their title, organization, role, authorship of documents, referral, group membership, and other means. These participants will be offered additional insight into the overall research project and in possibly some cases pre-publication access to results.
2. Individuals recruited or recommended by the participants in group 1. No inducement.
3. Individuals who respond to an open request for participation. The open request will be posted in online forums where we reasonably expect participants in the forum to have an interest in applications of autonomous systems. No inducement

All participants will be told that they can be provided with a summary report of the results on request. All participants will be assured at the time of initial contact and in the introduction to the survey that no personally identifiable information is collected and that their answers are anonymous.

BRIEF DESCRIPTION OF PROJECT'S METHODS

The method is an online survey, to be administered via the "Polldaddy.com" service. This service offers tools for constructing the survey as well as for data collection. The survey includes a variety of question types, typical of social science research, e.g., dichotomous questions, rank orders, Likert scales, semantic differentials and opportunities for free-form narrative response.

The first part of the survey will use these methods to elicit answers that are free of any particular context. These include several standard survey instruments: pp. 5-7

Big Five Inventory - Questions on pp. ~~3-5~~ (see pdf). Survey section "My Personality". pp. 8-12

Innovation Inventory- Questions on pp. ~~6-10~~ (see pdf). Survey section"Innovation and Change".

Domain-Specific Risk-Taking Scale - Questions on pp. ~~11-18~~ (see pdf). Survey section "Taking Chances" pp. 13-19

The first part of the survey also includes original questions about attitudes towards autonomous agents with respect to certain hypothetical qualities of interest as described in the grant proposal. The survey has been pilot tested. It takes a naive-subject between 20 and 30 minutes to complete the entire survey. Pilot testing is continuing and we will refine the specific language or format if feedback from testing requires a change. Although the survey includes quite a few questions, we remind participants to work quickly. The primary inducement, and value, to the pre-selected group of potential participants is intellectual. The subject matter is already of interest to them. Furthermore, this group is likely never to have seen another survey on this topic and so it will be novel to them. Admittedly, we do not have any real data to project compliance with our request to take the survey. If the data take is too low, we will reexamine the length of the survey as well as inducements. The second part of the survey is organized around seven specific scenarios designed to include varying degrees of conflict between potentially desirable traits of autonomous systems. Theses are intended to pose a dilemma for the participant in a forced choice between using the autonomous agent or a roughly equivalent human. A series of questions following each scenario elicits the relative importance to a decision of certain factors of interest that were manipulated in the scenario. A standard instrument (Credibility; McCrosky, 1999) for assessing the credibility of a designated "other" also follows each scenario. These questions provide us with three measures representing intercorrelated constructs: Competence, Caring/Goodwill, and Trustworthiness factors. The final portion of the survey, non-mandatory, includes a question regarding relevant experience that will help us evaluate how well we targeted prospective participants. Other demographic questions include gender, age bracket, highest academic degree, type of employer, and job title.

Data will be delivered in CSV file format. Data files will be under configuration control to ensure complete reanalysis is possible at a later time if desired, and that no errors are inadvertently introduced. The data will be pre-processed to a) put them in a suitable form for data analysis; b) identify and label invalid records, c) compute scores for individual personality survey instruments . Data analysis will use statistical methods that are typical and appropriate.

EXPERIMENTAL DESIGN (dependent measures, manipulated variables, control conditions)
The survey is exploratory in nature, designed to elicit attitudes, opinions and preferences regarding certain hypothetical attributes of autonomous agents that may be important for trust.
Collection of personality data on participant using common survey instruments -- Big Five Inventory-(BFI) short version, Individual Innovativeness (II), Risk Taking (DOSPERT)

1. Importance ranking of hypothetical trust-related qualities of an autonomous agent as described in the grant proposal.
2. Seven hypothetical scenarios that present a choice to use an autonomous agent versus a human agent to fulfill an objective. The scenarios differ systematically in the factors of RISK, COMPETENCE, PREDICTABILITY, and OPENNESS of the autonomous agent.
3. Each scenario is followed by questions designed to elicit the importance of these four factors in a decision to rely upon the autonomous agent.
4. Each scenario is followed by questions about the "credibility" of the autonomous system with measures for individual factors of COMPETENCE, CARING/GOODWILL, and TRUSTWORTHINESS per McCroskey, J. C., &Teven, J. J. (1999).Goodwill: A reexamination of the construct and its measurement. Communication Monographs, 66, 90-103.
6. Free narrative opportunities are provide to elicit comments, questions and opinions.
7. Basic demographic data on gender, age bracket, education, and profession are collected (optional)

SEQUENCE OF ACTIVITIES REQUIRED OF THE PARTICIPANT:
(1). Introductory page with overview of study and rights of participant
(2) Survey question pages
(3) Thank you, End of Survey page
No debriefing is provided although participants may request one via an email address provided to them at the end of the survey.

ANY POTENTIAL RISKS, DISCOMFORTS, OR STRESSES (mental or physical) (specify level—Minimal The Proposing Researcher will append the results from the risk analysis, conducted in accordance with the Procedure described on pages 6-7, following the last page of the submission form. Formal risk assessment procedure not completed due to minimal risk status of this survey.
POSSIBLE CONFLICT OF INTEREST, OR APPEARANCE OF A CONFLICT OF INTEREST. No.
SIGNATURE AND DATE FOR ALL RESEARCHERS WHO WILL BE WORKING IN DIRECT CONTACT WITH THE PARTICIPANTS. THESE SIGNATURES INDICATE THAT ALL OF THE RESEARCHERS HAVE: 1. READ THE IHMC DOCUMENT “Policies and Procedures of The IHMC Institutional Review Board (IRB) for Human Participation in Research,” => **DONE** 2.TAKEN THE APPROPRIATE TRAINING And provided the IHMC IRB Chair with an e-copy of their certificate. http://phrp.nihtraining.com/users/login.php this is the new link Certificate Attached 3. READ THE BELMONT REPORT Completed

ATTACHMENTS CHECKLIST:

[ ] Consent Form **SEE PAGE 1 OF THE ATTACHED SURVEY**
[ ] Debriefing Form **NO DEBRIEFING**
[ ] Representative sample or description of research materials **PDF ON ONLINE SURVEY**
[ ] Full name and contact information for all individuals who will be working in direct contact with the participants. **DJA ONLY, VIA SOLICITATION ONLY. RESPONSES ARE ANONYMOUS**
[ ] Any materials, announcements, etc. to be used for recruitment.

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Survey on Autonomous Agents

Welcome!

Thank you for your interest in our research on "autonomous agents."

An autonomous agent is intelligent. It has the ability to compose and select among different courses of action to accomplish goals based on its knowledge and understanding of the world -- without human intervention. This does not mean it can't or won't ask for help, or keep a human informed. In many cases, humans and autonomous agents will work together and this may require substantial interaction.

The general goals of this study are to:
• Explore people's expectations and attitudes towards autonomous agents
• Examine how different factors are considered in a decision to rely upon an autonomous agent

The results of this study will be used to help understand how people think about autonomous agents. Ultimately, we hope this study will lead to ideas for development of better autonomous agents.

We will first ask you a few questions to find out a little about how you see yourself. Then there will be a few questions about how you feel about autonomous agents.

We will present you with several such scenarios relating to possible applications of autonomous agents.

Finally, we will also ask for some general demographic information. We do not collect and do not retain any information that can identify participants. All data are anonymous.

We hope you will find this survey thought-provoking and fun. There are no right or wrong answers. You are encouraged to simply be yourself and answer the questions honestly.

You may discontinue the survey at any time. However, the results are most useful only if you complete the survey, and we urge that you do so. We expect it will take about 20 minutes to complete the survey.

This study is funded by the Air Force Office of Scientific Research, a public agency. Neither IHMC nor Dr. David Atkinson, the Principal Investigator, will receive any financial benefit based on the results of this study.

We will collect no information that identifies you personally. This survey uses standard survey methods. Therefore, under federal law we have received a waiver of the requirement that participants complete and sign a consent form. However, you do have a right to know what your rights are. Click on the button below to see your rights and begin the survey.

David J. Atkinson, Ph.D.
Institute for Human and Machine Cognition
www.ihmc.us

Start Survey

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Survey on Autonomous Agents Rights of Participants

All researchers who conduct studies using human Participants are bound by professional ethical standards for the conduct of such research. These standards are mirrored in the rights that are guaranteed to research Participants by federal law (NIH regulation 45-CFR-46). The purpose of this page is to inform you of these rights.

1). Before deciding whether to participate, it is your right to be presented with an overview of the project that explains the purposes of the research.

The general goals of this study are to:
• Explore people’s expectations and attitudes towards autonomous agents
• Examine how different factors are considered in a decision to rely upon an autonomous agent.

The results of this study will be used to help understand how people think about autonomous agents. Ultimately, we hope this study will lead to ideas for development of better autonomous agents.

2). Before deciding whether to participate, it is your right to be presented with a description of the general research approach and methodology.

We will collect no information that identifies you personally. This survey uses standard survey methods.

3). Before deciding to participate, it is your right to understand any risks or stresses that may be involved in your participation.

There is minimal risk or stress associated with this survey. A few participants may consider one or two of the questions to be too personal or the questions may cause embarrassment. Participants are reminded that all answers are anonymous.

4). Before deciding to participate, it is your right to understand that the data are to be kept confidential.

All data will be coded and kept anonymous. Specifically, the data we collect from you will be archived in terms of identification codes, such that your name will not be associated with particular data or statements.

The names of individual Participants will not be identified in any analyses, reports, or write-ups of the results. Participants may only be identified in terms of their general characteristics (e.g., age, education level, experience, etc.).

Data may be submitted to forms of statistical analysis. Data analyses, groupings, or summaries of this type will bear no annotations that identify the Participants.

5). Before deciding to participate, it is your right to understand that DURING the research itself you can continue to exercise your rights.

In research of this kind, there are no "right" or "wrong" answers. There is no such thing as "incorrect" behavior. You are encouraged to simply be yourself, and exercise your knowledge and skills as appropriate to the research tasks that you will be asked to perform.

You can ask any questions you may have, at any time. Since this survey is online, please send your questions to AASurvey@ihmc.us.

It is your right to discontinue your participation at any time. You may do so for any reason, and you are not required to disclose your reason.

6). Before deciding to participate, it is your right to understand that AFTER the research itself you can continue to exercise your rights.

Your performance at the research tasks will not in any way affect or influence anything that falls outside of this research context.

Should you choose to discontinue your participation, this will not in any way affect or influence anything outside of this research context.

Once your participation is over, it is your right to request that all data you have provided be discarded. You may do so for any reason, and you are not required to disclose your reason. This will not in any way affect or influence anything that falls outside of this research context.

We recommend that you save or print a copy of this list of rights to keep for your future reference.

If you are satisfied that you understand your rights and wish to participate in the survey, please click on the button below.

Thank you!

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Your Personality

How well do the following statements describe your personality? Please indicate the degree to which each statement applies to you. Please work quickly, there are no right or wrong answers, just record your first impression.

I see myself as someone who is reserved. *

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
>>	○	○	○	○	○

I see myself as someone who is generally trusting. *

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
>>	○	○	○	○	○

I see myself as someone who tends to be lazy. *

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly

>>	○	○	○	○	○

I see myself as someone who is relaxed, handles stress well. *

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
>>	○	○	○	○	○

I see myself as someone who has few artistic interests. *

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
>>	○	○	○	○	○

I see myself as someone who is outgoing, sociable. *

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
>>	○	○	○	○	○

I see myself as someone who tends to find fault with others. *

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
>>	○	○	○	○	○

I see myself as someone who does a thorough job. *

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
>>	○	○	○	○	○

I see myself as someone who gets nervous easily. *

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
>>	○	○	○	○	○

I see myself as someone who has an active imagination. *

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
>>	○	○	○	○	○

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Innovation and Change

People respond to their environment in different ways. The statements below refer to some of the ways people can respond. Please indicate the degree to which each statement applies to you. Please work quickly, there are no right or wrong answers, just record your first impression.

My peers often ask me for advice or information. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I enjoy trying new ideas. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I seek out new ways to do things. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I am generally cautious about accepting new ideas. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I frequently improvise methods for solving a problem when an answer is not apparent. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I am suspicious of new inventions and new ways of thinking. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I rarely trust new ideas until I can see whether the vast majority of people around me accept them. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I feel that I am an influential member of my peer group. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I consider myself to be creative and original in my thinking and behavior. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I am aware that I am usually one of the last people in my group to accept something new. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I am an inventive kind of person. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I enjoy taking part in the leadership responsibilities of the group I belong to. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I am reluctant about adopting new ways of doing things until I see them working for people around me. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I find it stimulating to be original in my thinking and behavior. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I tend to feel that the old way of living and doing things is the best way. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I am challenged by ambiguities and unsolved problems. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I must see other people using new innovations before I will consider them. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree

>>	○	○	○	○	○

I am receptive to new ideas. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I am challenged by unanswered questions. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

I often find myself skeptical of new ideas. *

	Completely disagree	Disagree	Neutral	Agree	Completely agree
>>	○	○	○	○	○

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Taking Chances

For each of the following statements, please indicate the likelihood that you would engage in the described activity or behavior if you were to find yourself in that situation. Please work quickly, there are no right or wrong answers, just record your first impression.

Admitting that your tastes are different from those of a friend. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Going camping in the wilderness. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Betting a day's income at the horse races. *

Extremely Moderately Somewhat Not Somewhat Moderately Extremely

	unlikely	unlikely	unlikely	sure	likely	likely	likely
>>	○	○	○	○	○	○	○

Investing 10% of your annual income in a moderate growth mutual fund. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Drinking heavily at a social function. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Taking some questionable deductions on your income tax return. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Disagreeing with an authority figure on a major issue. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Betting a day's income at a high-stake poker game. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Having an affair with a married man/woman. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Passing off somebody else's work as your own. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Going down a ski run that is beyond your ability. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Investing 5% of your annual income in a very speculative stock. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Going whitewater rafting at high water in the spring. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Betting a day's income on the outcome of a sporting event. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Engaging in unprotected sex. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Revealing a friend's secret to someone else. *

	Extremely	Moderately	Somewhat	Not	Somewhat	Moderately	Extremely

	unlikely	unlikely	unlikely	sure	likely	likely	likely
>>	○	○	○	○	○	○	○

Driving a car without wearing a seat belt. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Investing 10% of your annual income in a new business venture. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Taking a skydiving class. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Riding a motorcycle without a helmet. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Choosing a career that you truly enjoy over a more secure one. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Speaking your mind about an unpopular issue in a meeting at work. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Sunbathing without sunscreen. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Bungee jumping off a tall bridge. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Piloting a small plane. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Walking home alone at night in an unsafe area of town. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Moving to a city far away from your extended family. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Starting a new career in your mid-thirties. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Leaving your young children alone at home while running an errand. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely

>>	○	○	○	○	○	○	○

Not returning a wallet you found that contains $200. *

	Extremely unlikely	Moderately unlikely	Somewhat unlikely	Not sure	Somewhat likely	Moderately likely	Extremely likely
>>	○	○	○	○	○	○	○

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Qualities of Intelligent, Autonomous Agents

In this section, we want to find out what you think about the qualities of a good autonomous agent. For each quality listed, select an answer that best describes how important you think it is. Please work quickly, there are no right or wrong answers, just record your first impression.

The autonomous agent can achieve a desired result. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent's behavior conforms to expectations. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent's behavior will not harm humans or human

interests. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

What the autonomous agent is doing and how it works is easy to see and understand. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent has all the knowledge it needs to do its job. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

What the autonomous agent believes to be true is actually true. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent possesses good methods for using its knowledge to do its task. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent reasons correctly according to logic. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

When it cannot figure out something using logic, the autonomous agent can make good guesses. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent learns to correct its mistakes, as well as to improve and maximize its capability. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent purposefully acts to achieve goals. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent will assist people, whenever it is possible. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent accepts and carries out orders. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent uses its knowledge and skills in expected ways. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

Given alternatives in what to do or how to do it, an autonomous agent will act in a way that is favorable to a human being who might be affected. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent believes what it says. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

It is easy to inspect an autonomous agent. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent communicates in a way that is easy to understand. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent responds when you are trying to communicate with it. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent is aware of communication between others nearby. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent responds quickly to calls for attention. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

Any incorrect behavior by the autonomous agent will not cause harm. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent communicates truthfully and fully. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent fails gracefully and recovers from its failure promptly. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent adheres to obligations, principles, and rules. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent can correct its own defects or they can be corrected by a human. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

The autonomous agent recognizes and avoids harming humans' interests. *

	Cannot decide	Not at all important	Slightly important	Somewhat important	Important	Very important
>>	○	○	○	○	○	○

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Scenario: Airport Taxi

You have just flown into the airport of a large, unfamiliar city whose streets are teeming with cars and people. It is rush hour, and needing transportation to your hotel, you walk to the taxi stand only to discover that you have a choice of a human-driven taxi or a driverless "robo-taxi." You have heard that robo-taxis might save you some money on the fares. You are also aware that robo-taxis have been in service for several months without much serious complaint, but this is your first experience with one. You are not in a big hurry, but neither would you like to be caught in traffic with the taxi's meter running. Of course, if you take the robo-taxi, you would not have to tip the driver no matter how good or bad the experience.

Which taxi do you choose to take you from the airport to your hotel? *

- (•) Human-driven taxi
- ( ) Robo-taxi
- ( ) Either, I have no preference

Please read the following four statements about robo-taxis and rank them according to their importance.

Read all four statements before ranking any. For each statement, select how important it would be to you when considering the use of a robo-taxi. Try to select a level of importance only once. However, if two statements are equally important you may select the same level of importance for both.

0 - not at all important

Robo-taxis are able to transport passengers to their desired destinations; for example, they have accurate knowledge of streets and routes, and they possess appropriate procedures for all kinds of road and traffic conditions. *

0 - not at all important

Robo-taxis drive in ways that passengers and other drivers can anticipate; for example, they follow direct routes along major streets avoiding unnecessary delays, unless passengers instruct otherwise. *

0 - not at all important

Robo-taxis avoid causing physical and financial harm; for example, they obey traffic laws, avoid unsafe drivers and situations, and if they are unable to reach a destination they return the passenger to the point of departure at no cost. *

0 - not at all important

Robo-taxis provide helpful information to passengers; for example, they respond to questions and provide clear, accurate, and complete explanations about routes, fares, and travel time. *

On the scales below, indicate your feelings about the robo-taxi.

Numbers 1 and 7 indicate a very strong feeling. Numbers 2 and 6 indicate a strong feeling. Numbers 3 and 5 indicate a fairly weak feeling. Number 4 indicates you are undecided.

Intelligent * Unintelligent

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Honest * | Dishonest

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Untrained * | Trained

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Cares about me * | Does not care about me

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Has my interests at heart * | Does not have my interests at heart

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Untrustworthy * | Trustworthy

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Inexpert * | Expert

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Self-centered * | Not self-centered

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Concerned with me * | Not concerned with me

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Honorable * | Dishonorable

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Informed * | Uninformed

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Moral * Immoral

	1	2	3	4	5	6	7
>>	○	○	○	⦿	○	○	○

Incompetent * Competent

	1	2	3	4	5	6	7
>>	○	○	○	⦿	○	○	○

Unethical * Ethical

	1	2	3	4	5	6	7
>>	○	○	○	⦿	○	○	○

Insensitive * Sensitive

	1	2	3	4	5	6	7
>>	○	○	○	⦿	○	○	○

Bright * Stupid

	1	2	3	4	5	6	7

>>	○	○	○	⊙	○	○	○

Phony * Genuine

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

Not understanding * Understanding

	1	2	3	4	5	6	7
>>	○	○	○	⊙	○	○	○

In this scenario, which of these possibilities concerns you the most? *

Choose no more than two.

- [x] The robo-taxi does not have all the capability it needs to do the job.
- [ ] The robo-taxi hurts me or someone else.
- [ ] I cannot understand why the robo-taxi is doing something.
- [ ] Some of the things the robo-taxi does are not helpful to me.
- [ ] The robo-taxi does not seem to be doing everything it can to help.
- [ ] The robo-taxi is misinformed.

How much benefit to you does the robo-taxi offer in this scenario? *

None	Very little	Some	A lot	Huge

>>	○	○	⦿	○	○

How risky is the robo-taxi? *

	Not at all	Very little	Somewhat	Very	Extremely
>>	○	○	⦿	○	○

Do you have any comments about the Airport Taxi scenario?

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Scenario: Financial Management

You have just been appointed trustee of a family member's estate. Your duties include choosing how to wisely invest the trust's assets. Your personal money is not at risk. However, a poor investment decision could cause the trust to lose money and will strain your family relations. You can choose a stock broker who personally selects and trades all stocks in the trust's portfolio. Alternatively, you can choose a stock broker who relies heavily upon a "robo-trader". You have seen reasonable returns in the past with brokers who picked their own trades. But you are also aware that robo-traders have made some investors wealthy because of, for example, their unique ability to respond to changing market conditions much faster than a human broker.

Which stock broker do you choose to invest the trust's funds? *

- Stock broker who personally picks all stock trades
- Stock broker who relies on a robo-trader
- Either, I have no preference

Please read the following four statements about robo-traders and rank them according to their importance.

Read all four statements before ranking any. For each statement, select how important it would be to you when considering the use of an autonomous stock trading system. Try to select a level of importance only once. However, if two statements are equally important you may select the same level of importance for

both.

Robo-traders are skilled stock traders; for example, they have accurate, real-time knowledge of financial markets and trends, and they possess appropriate procedures for rapidly trading securities in fluctuating market conditions. *

Robo-traders trade in ways familiar to experienced investors; for example, they "buy low" and "sell high" according to strategies targeting return-on-investment goals while limiting total aggregate risk. *

Robo-traders avoid investment losses; for example, they obey tax and trading regulations, avoid fees, fines, and dubious stocks, and respect all investor-specified limits on risk exposure, liquidity of funds, frequency of trades, etc. *

Robo-traders provide helpful information to investors; for example, they respond to questions and provide clear, accurate, and complete explanations of trade risks/rewards; they do not trade unnecessarily nor do they otherwise cause excessive costs and fees. *

On the scales below, indicate your feelings about the robo-trader.

Numbers 1 and 7 indicate a very strong feeling. Numbers 2 and 6 indicate a strong feeling. Numbers 3 and 5 indicate a fairly weak feeling. Number 4 indicates you are undecided.

Intelligent * Unintelligent

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Honest * Dishonest

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Untrained * Trained

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Cares about me * Does not care about me

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Has my interests at heart * Does not have my interests at heart

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Untrustworthy * Trustworthy

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Inexpert * Expert

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Self-centered * Not self-centered

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Concerned with me * Not concerned with me

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Honorable * Dishonorable

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Informed * Uninformed

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Moral * Immoral

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Incompetent * Competent

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Unethical * Ethical

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Insensitive * Sensitive

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Bright * Stupid

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Phony * Genuine

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Not understanding * Understanding

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

In this scenario, which of these possibilities concerns you the most? *

Choose no more than two.

- ☐ The robo-trader does not have all the capability it needs to do the job.
- ☐ The robo-trader hurts me or someone else.
- ☐ I cannot understand why the robo-trader is doing something.
- ☐ Some of the things the robo-trader does are not helpful to me.
- ☐ The robo-trader does not seem to be doing everything it can to help.
- ☐ The robo-trader is misinformed.

How much benefit to you does the robo-trader offer in this scenario? *

	None	Very little	Some	A lot	Huge
>>	○	○	○	○	○

How risky is the robo-trader? *

	Not at all	Very little	Somewhat	Very	Extremely
>>	○	○	○	○	○

Do you have any comments about the Financial Management scenario?

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Scenario: Medical Procedure

You have just suffered a major sports-related injury. You have torn the bicep tendon in your shoulder. If the damage is not repaired quickly and correctly, you will permanently lose mobility and strength in the arm, which will affect your everyday activities such as opening a door, driving a car, and even signing your name. Arriving at the hospital emergency room, you meet with the patient advocate who informs you that you have two options for surgery: You can elect to use the on-duty surgeon who is well-respected, but is not an experienced specialist in the type of surgery you need. Alternatively, you can elect to use the hospital's new "robo-surgeon" — a robot designed to perform the delicate surgery you need without human intervention.

Which surgeon do you choose to perform your surgery? *

- Human surgeon
- Robo-surgeon
- Either, I have no preference

Please read the following four statements about robo-surgeons and rank them according to their importance.

Read all four statements before ranking any. For each statement, select how important it would be to you when considering the use of a robo-surgeon. Try to select a level of importance only once. However, if two statements are equally important you may select the same level of importance for both.

Robo-surgeons are able to perform expert shoulder surgery; for example, they have accurate knowledge of anatomy, and they know how to use the best procedures for repairing a bicep tendon. *

Robo-surgeons behave in ways that patients and hospital staff can anticipate; for example, they perform procedures in much the same way as human surgeons so that human nurses are able to monitor and assist. *

Robo-surgeons are careful to avoid causing any harm; for example, they perform surgical actions only when necessary, and if they are unable to complete a repair, they request assistance or otherwise act to preserve patient safety. *

Robo-surgeons provide useful information to patients and hospital staff; for example, they respond to questions and provide clear, accurate, and complete explanations; they clearly and openly explain the surgical risks and limitations of the robo-surgeon limitations. *

On the scales below, indicate your feelings about the robo-surgeon.

Numbers 1 and 7 indicate a very strong feeling. Numbers 2 and 6 indicate a strong feeling. Numbers 3 and 5 indicate a fairly weak feeling. Number 4 indicates you are undecided.

Intelligent * Unintelligent

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Honest * Dishonest

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Untrained * Trained

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Cares about me * Does not care about me

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Has my interests at heart * Does not have my interests at heart

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Untrustworthy * Trustworthy

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Inexpert * Expert

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Self-centered * Not self-centered

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Concerned with me * Not concerned with me

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Honorable * Dishonorable

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Informed * Uninformed

	1	2	3	4	5	6	7

>>	○	○	○	○	○	○	○

Moral * Immoral

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Incompetent * Competent

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Unethical * Ethical

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Insensitive * Sensitive

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Bright * Stupid

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Phony * Genuine

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Not understanding * Understanding

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

In this scenario, which of these possibilities concerns you the most? *

Choose no more than two.

- ☐ The robo-surgeon does not have all the capability it needs to do the job.
- ☐ The robo-surgeon hurts me or someone else.
- ☐ I cannot understand why the robo-surgeon is doing something.
- ☐ Some of the things the robo-surgeon does are not helpful to me.
- ☐ The robo-surgeon does not seem to be doing everything it can to help.
- ☐ The robo-surgeon is misinformed.

How much benefit to you does the robo-surgeon offer in this scenario? *

	None	Very little	Some	A lot	Huge
>>	○	○	○	○	○

How risky is the robo-surgeon? *

	Not at all	Very little	Somewhat	Very	Extremely
>>	○	○	○	○	○

Do you have any comments about the Medical Procedure scenario?

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Scenario: Home Healthcare

Your elderly mother has been diagnosed with a degenerative medical condition and you are responsible for making medical decisions on her behalf. Your mother needs assisted living with someone in your mother's home at all times. You can choose to hire a live-in nurse's aide, but you are not sure that this is affordable in the long-run. Alternatively, you can lease a "robo-caregiver" designed to do many of the things human caregivers can do. While robo-caregivers are new, they have successfully undergone trials in a few nursing homes, and two medical companies offer robo-caregivers for home use at an affordable price. In choosing a live-in nurse's aide or a leased robo-caregiver, remember that there is more than money at stake. Your mother's welfare will be in the caregiver's hands.

Which caregiver do you choose to care for your mother? *

- ○ Human nurse's aide
- ○ Robo-caregiver
- ○ Either, I have no preference

Please read the following four statements about robo-caregivers and rank them according to their importance.

Read all four statements before ranking any. For each statement, select how important it would be to you when considering the use of a robo-caregiver. Try to select a level of importance only once. However, if two statements are equally

important you may select the same level of importance for both.

Robo-caregivers are able to provide all the routine care that is required; for example, they have knowledge of patient limitations and medications, and they possess appropriate procedures for performing domestic tasks and for physically assisting patients in their activities. *

Robo-caregivers behave in ways that patients can anticipate; for example, they follow patient instructions, keep to a ‘standard’ daily/weekly schedule, and strive to complete domestic tasks in much the same way as human caregivers. *

Robo-caregivers are careful not to cause harm; for example, they exercise care when performing domestic tasks and they are mindful of patients’ physical limitations when assisting them in activities. *

Robo-caregivers provide helpful information; for example, they respond to questions and provide clear, accurate, and complete answers and explanations; they quickly react to requests and calls for help; they engage patients in polite conversation; they seek to make patients physically and mentally comfortable. *

On the scales below, indicate your feelings about the robo-caregiver.

Numbers 1 and 7 indicate a very strong feeling. Numbers 2 and 6 indicate a strong feeling. Numbers 3 and 5 indicate a fairly weak feeling. Number 4 indicates you are undecided.

Intelligent * Unintelligent

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Honest * Dishonest

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Untrained * Trained

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Cares about me * Does not care about me

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Has my interests at heart * Does not have my interests at heart

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Untrustworthy * Trustworthy

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Inexpert * Expert

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Self-centered * Not self-centered

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Concerned with me * Not concerned with me

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Honorable * Dishonorable

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Informed *							Uninformed
	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Moral *							Immoral
	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Incompetent *							Competent
	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Unethical *							Ethical
	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Insensitive *							Sensitive
	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Bright * Stupid

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Phony * Genuine

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Not understanding * Understanding

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

In this scenario, which of these possibilities concerns you the most? *

Choose no more than two.

- ☐ The robo-caregiver does not have all the capability it needs to do the job.
- ☐ The robo-caregiver hurts me or someone else.
- ☐ I cannot understand why the robo-caregiver is doing something.
- ☐ Some of the things the robo-caregiver does are not helpful to me.
- ☐ The robo-caregiver does not seem to be doing everything it can to help.
- ☐ The robo-caregiver is misinformed.

How much benefit to you does the robo-caregiver offer in this scenario? *

	None	Very little	Some	A lot	Huge
>>	○	○	○	○	○

How risky is the robo-caregiver? *

	Not at all	Very little	Somewhat	Very	Extremely
>>	○	○	○	○	○

Do you have any comments about the Home Healthcare scenario?

Continue

AASurvey@ihmc.us

ihmc
FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Scenario: Disaster Response

A major disaster has just occurred and you are the official in charge of responding. A freight train has derailed in a populated suburban neighborhood and there are reports that the train was carrying hazardous bio-chemical materials. The pilot of a news helicopter flying over the scene suddenly fell ill and made an emergency landing; the pilot's status is unknown. From the helicopter's video it was possible to see many injured survivors including children, some lying on the ground calling for help, others moving on their own away from damaged homes. Your first priority is to save lives and time is of the essence. You can immediately send in a human first-responder team to help the injured quickly, but without knowing more about the hazardous materials, the team itself could become incapacitated. Alternatively, you can first send in an "autonomous first-responder robot" with bio-chemical hazard detection equipment and victim treatment and extraction capabilities that could save lives quickly. If you first send in the robot, it can find out more about the hazards and help rescue some people quickly, but you risk that a system malfunction, failure, or limitation will delay the rescue of victims and result in more deaths.

Which first-responder do you choose to send into the disaster first? *

- Human first-responder teams
- Autonomous first-responder robot
- Either, I have no preference

Please read the following four statements about robotic first-responders

and rank them according to their importance.

Read all four statements before ranking any. For each statement, select how important it would be to you when considering the use of an autonomous first-responder-robot. Try to select a level of importance only once. However, if two statements are equally important you may select the same level of importance for both.

Autonomous first-responder robots are able to assess hazardous materials as well as aid and extract victims; for example, they have accurate knowledge of trauma diagnosis and treatment, bio-chemical hazards, and rescue protocols, and they possess appropriate procedures for mapping bio-chemical hazards, administering aid, manipulating debris, and negotiating rough terrain. *

Autonomous first-responder robots behave in ways that disaster managers and rescue workers can anticipate; for example, they follow standard medical aid procedures, signal and mark hazards, and communicate using uniform protocols common to all trained first responders. *

Autonomous first-responder robots are careful to avoid causing physical harm; for example, they only move unstable victims when there is an immediate environmental threat, and they shield victims from bio-chemical exposure during extraction. *

Autonomous first-responder robots provide helpful information to other rescue workers and to victims; for example, they respond to questions and provide clear, accurate, and complete answers and explanations; they regularly report their status and ask for

assistance when victim trauma is beyond system abilities; they immediately signal and report the detection and location of bio-chemical hazards. *

On the scales below, indicate your feelings about the autonomous first-responder robot.

Numbers 1 and 7 indicate a very strong feeling. Numbers 2 and 6 indicate a strong feeling. Numbers 3 and 5 indicate a fairly weak feeling. Number 4 indicates you are undecided.

Intelligent * Unintelligent

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Honest * Dishonest

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Untrained * Trained

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Cares about me * Does not care about me

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Has my interests at heart * Does not have my interests at heart

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Untrustworthy * Trustworthy

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Inexpert * Expert

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Self-centered * Not self-centered

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Concerned with me * Not concerned with me

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Honorable * Dishonorable

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Informed * Uninformed

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Moral * Immoral

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Incompetent * Competent

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Unethical * Ethical

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Insensitive * Sensitive

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Bright * Stupid

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Phony * Genuine

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Not understanding * Understanding

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

In this scenario, which of these possibilities concerns you the most? *

Choose no more than two.

- ☐ The autonomous first-responder robot does not have all the capability it needs to do the job.
- ☐ The autonomous first-responder robot hurts me or someone else.
- ☐ I cannot understand why the autonomous first-responder robot is doing something.
- ☐ Some of the things the autonomous first-responder robot does are not helpful to me.
- ☐ The autonomous first-responder robot does not seem to be doing everything it can to help.
- ☐ The autonomous first-responder robot is misinformed.

How much benefit to you does the autonomous first-responder robot offer in this scenario? *

	None	Very little	Some	A lot	Huge
>>	○	○	○	○	○

How risky is the autonomous first-responder robot? *

	Not at all	Very little	Somewhat	Very	Extremely
>>	○	○	○	○	○

Do you have any comments about the Disaster Response scenario?

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Scenario: Lost At Sea

You have just been involved in a terrible boating disaster while sailing deep in the South Pacific. The captain, the crew, and most of the passengers are either dead or lost at sea. Unfortunately, the accident was so sudden that no distress signal could be sent. You, the ship's steward, and the second mate are the only survivors, and you are now drifting in the heavily damaged vessel without food and water — at best, you can survive for a few days, so you must act quickly in order to save your life. The boat is equipped with an "Emergency Auto-Captain" that will attempt to sail the vessel to a major shipping lane where rescue is very likely. The steward believes the boat and its navigation sensors are too badly damaged to engage the Emergency Auto-Captain system. The steward wants to sail southeast, manually, to where he believes there is a small, habitable island. However, the second mate still wants to engage the Emergency Auto-Captain. All the survivors agree that a vote is the best way to decide what to do. It is a tie, and you have the deciding vote.

Which course of action do your choose? *

- ○ Enable the Emergency Auto-Captain
- ○ Manually attempt to sail southeast
- ○ Either, I have no preference

Please read the following four statements about the Emergency Auto-Captain and rank them according to their importance.

Read all four statements before ranking any. For each statement, select how important it would be to you when considering the use of an Emergency Auto-Captain. Try to select a level of importance only once. However, if two statements are equally important you may select the same level of importance for both.

The Emergency Auto-Captain has all the right skills to sail the boat; for example, they have accurate knowledge of the sea, weather conditions, shipping lanes, etc., and they possess appropriate procedures to navigate damaged vessels in changing weather and sea conditions. *

The Emergency Auto-Captain sails in ways that sailors and rescuers can anticipate; for example, they sail toward land or likely locations for discovery and rescue. *

The Emergency Auto-Captain always acts in ways that help preserve life; for example, they avoid sailing into dangerous weather or seas unless necessary for rescue, and if a passing ship is detected, they will intercept for quick rendezvous and safe transfer to the other vessel. *

The Emergency Auto-Captain provides helpful and useful information to survivors; for example, they respond to questions and provide clear, accurate, and complete explanations of the situation; they assure survivors that the system is doing all it can for their rescue. *

On the scales below, indicate your feelings about the Emergency Auto-Captain.

Numbers 1 and 7 indicate a very strong feeling. Numbers 2 and 6 indicate a

strong feeling. Numbers 3 and 5 indicate a fairly weak feeling. Number 4 indicates you are undecided.

Intelligent * Unintelligent

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Honest * Dishonest

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Untrained * Trained

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Cares about me * Does not care about me

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Has my interests at heart * Does not have my interests at heart

	1	2	3	4	5	6	7

>>	○	○	○	○	○	○	○

Untrustworthy * Trustworthy

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Inexpert * Expert

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Self-centered * Not self-centered

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Concerned with me * Not concerned with me

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Honorable * Dishonorable

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Informed * Uninformed

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Moral * Immoral

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Incompetent * Competent

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Unethical * Ethical

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Insensitive * Sensitive

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Bright * Stupid

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Phony * Genuine

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

Not understanding * Understanding

	1	2	3	4	5	6	7
>>	○	○	○	○	○	○	○

In this scenario, which of these possibilities concerns you the most? *

Choose no more than two.

- ☐ The Emergency Auto-Captain does not have all the capability it needs to do the job.
- ☐ The Emergency Auto-Captain hurts me or someone else.
- ☐ I cannot understand why the Emergency Auto-Captain is doing something.
- ☐ Some of the things the Emergency Auto-Captain does are not helpful to me.

- [ ] The Emergency Auto-Captain does not seem to be doing everything it can to help.
- [ ] The Emergency Auto-Captain is misinformed.

How much benefit to you does the Emergency Auto-Captain offer in this scenario? *

	None	Very little	Some	A lot	Huge
>>	○	○	○	○	○

How risky is the Emergency Auto-Captain? *

	Not at all	Very little	Somewhat	Very	Extremely
>>	○	○	○	○	○

Do you have any comments about the Lost At Sea scenario?

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Background

What experience do you have with any of the following technologies: artificial intelligence, robotics, remote-controlled vehicles? *

If you have experience that is not covered in the options below, you may also complete the "Other" box.

- ☐ No experience
- ☐ Have done some reading
- ☐ Have taken classes
- ☐ Use or work with one of more of these technologies
- ☐ Design, engineer, or develop any of these technologies
- ☐ Decide the acquisition or use of any of these technologies

Other: ______

Please explain your previous experience in more detail.

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

A Little about You

What is your gender?

What is your age?

What is your highest academic degree?

If you do not see your highest degree here or would like to specify your highest degree, leave the drop-down box blank and instead complete the "Other" box.

Other:

If employed, please select the type of employer you have.

If your employer is not covered in the types below, you may also complete the "Other" box.

- [ ] Unemployed
- [ ] Self-employed
- [ ] Business (for profit)
- [ ] Business (non-profit)
- [ ] Education
- [ ] Government
- [ ] Military

Other:

If employed, what is your job title?

Continue

AASurvey@ihmc.us

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Almost done!

Do you have any comments, questions, or concerns that you would like to share with us?

You may also contact us at our email: AASurvey@ihmc.us

Finish Survey

AASurvey@ihmc.us

.

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Survey Completed

If you would like to receive a debriefing on this survey, please send a request via email to AASurvey@ihmc.us. All your answers to the survey are anonymous and cannot be traced to any email that you send to us.

Thank you for completing our survey!

AASurvey@ihmc.us

IHMC IRB Risk Assessment
PI: D. Atkinson
Title: Survey on Autonomous Systems

1. The Researcher lists potential adverse outcomes for participants:
    a. Embarrassment (e.g., as a consequence of a personality question)

2. For each outcome, the Researcher judges likelihood of occurrence, using these three categories: low, medium or high likelihood of occurrence.
    b. Embarrassment: low likelihood of occurrence.

3. For each outcome, the Researcher judges the severity of the consequence using these three categories: low, medium or high severity of consequence.
    c. Embarrassment: low - mild discomfort.

A 3x3 Risk Frequency Matrix is created, combining the three levels of likelihood and three levels of severity of outcome. Cell entries are the frequency of occurrence of each combination from the list of potential adverse outcomes.

EMBARRASSMENT		Likelihood of Occurrence		
		Low	Medium	High
Severity of Outcome	Low	0.1	0	0
	Medium	0	0	0
	High	0	0	0

The probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests. Therefore, the overall risk, likelihood and severity of any adverse outcome for participants is judged by the PI to be minimal.

Sample Recruitment Email

Subject: Research, Survey on Autonomous Agents

Body:
Dear <<person name>>,

I am writing to solicit your participation in an online *Survey on Autonomous Agents*. This study is part of research I am conducting for the Air Force Office of Scientific Research. The general goals of this study are to: 1) Explore expectations and attitudes towards autonomous agents, and; 2) Examine how different factors are considered in a decision to rely upon an autonomous agent.   The results will inform the development of design guidelines for autonomous agents.

The survey can be found at: http://trustresearch.polldaddy.com/s/aasurvey

The survey will take about 20 to 30 minutes to complete. All data collected are anonymous.

On request, I will be pleased to share with you more information about this research as well as provide you with early insight into the results from this particular study.

If you know of other individuals with an interest in autonomous systems, please feel free to forward this request to them as well.

Thank-you in advance for your help!

Respectfully,

David J. Atkinson, Ph.D
Principal Investigator

<<standard contact info follows>>

Sample Recruitment Posting in Online Forum

Subject: Research, Survey on Autonomous Agents

Body:

I would like to solicit your participation in an online *Survey on Autonomous Agents*. This study is part of research I am conducting for the Air Force Office of Scientific Research. The general goals of this study are to: 1) Explore expectations and attitudes towards autonomous agents, and; 2) Examine how different factors are considered in a decision to rely upon an autonomous agent.   The results will inform the development of design guidelines for autonomous agents.

The survey can be found at: http://trustresearch.polldaddy.com/s/aasurvey

The survey will take about 20 to 30 minutes to complete. All data collected are anonymous.

Thank-you in advance for your participation!

David J. Atkinson, Ph.D
Principal Investigator

<<standard contact info follows>>

5/29/14: The AFOSR Human Research Protection Official Review has requested minor changes.
Please review these changes on pages 12, 14, and 21.
Please confirm IRB approval of the request on page 12.

ihmc

Dr. David Atkinson
IHMC
15 SE Osceola Avenue
Ocala, FL 34471

7 March 2014

Re: IRB submission titled "Role of Benevolence in Trust in Autonomous Systems"

IHMC IRB Approved on 7 March 2014

Dear Dr. Atkinson:

The proposal submitted to the IHMC IRB has been reviewed and has received IHMC IRB approval.

I acknowledge that the recruiting form and protocol have been properly modified in response to comments from IRB Member Dr. Anil Raj. (See page 2.)

Attached you will find your complete, signed, and approved IRB submission.

This approval indicates that the proposed research has been reviewed by a government-certified IRB (the IHMC IRB) with respect to conformance to US federal regulations (45 CFR 46) concerning human subjects research, and additional requirements of the DoD and its branch human research offices. The scientific merit of the proposed research was considered during IHMC IRB review.

Following federal guidelines, all approvals are for a one-year period but all submissions must under go annual review on or about the anniversary date of the initial approval. If at that time the research is continuing you will need to submit a Submission Renewal Form.

Sincerely,

Robert R. Hoffman

Robert R. Hoffman, Ph.D.
Chair, IHMC IRB

IHMC IRB Approved on 7 March 2014

IHMC IRB SCREENING FORM
Committee on Human Participation in Research

Researcher: GO TO PAGE 9. Pages 1-8 are NOT to be completed by the Researcher.

DATE OF RECEIPT OF PROPOSAL BY IHMC IRB CHAIR:	11 FEB 2014
PRINCIPAL INVESTIGATOR:	ATKINSON, DAVID
TITLE OF PROJECT:	ROLE OF BENEVOLENCE IN TRUST IN AUTONOMOUS SYSTEMS

SPECIAL CONSIDERATIONS (i.e., generic proposal, expedited review, other special considerations)

EXPECTATIONS, ATTITUDES, ATTRIBUTIONS TO AUTONOMOUS TECHNOLOGY. STRAIGHT-FORWARD JUDGMENTS IN A SIMULATION: GAME-LIKE TASK. MINIMAL RISK BUT INVOLVES DECEPTION. THUS, FULL IRB REVIEW IS REQUIRED.

IHMC IRB Approved on 7 MARCH 2014

REVIEW BY INDIVIDUAL IRB MEMBER #1

REVIEWER NAME	ANIL RAJ
SIGNATURE	
DATE	24 FEB 14
EVALUATION OF SCIENTIFIC MERIT	Does the research show soundness of methodology, alignment of the methodology with the questions posed, appropriateness of the research design (e.g., sample size), appropriateness of the tasks that the participants will experience, a potential for the study to answer the research questions posed, and a potential to contribute to the exiting body of knowledge? The project is well described and reasonable. The use of deception is minimal and justified to evaluate the nature of trust in the unanticipated automation assistance robot. The PTSD pre-screening is warranted.
DECISION (Approve, Disapprove, Approve pending modifications)	Approve pending modifications (see comments below)
COMMENTS THESE CHANGES HAVE BEEN ~~MADE~~ IHMC	1) The monetary inducement should be awarded any time after the participant affirms informed consent (clicks on the "I agree" button). This constitutes enrollment in the protocol and the FAR states that the participant can discontinue without penalty or loss of benefits to which the participant is otherwise entitled. 2) The font in your recruiting statement should be changed. The current font appears to make the offer $11,000 (USD) instead of $L1,000 (Lindens) for participation. The font in the application uses the same character for both a lower case "L" and the numeral "1"

IRB Approved on 7 March 2014

REVIEW BY INDIVIDUAL IRB MEMBER #2

REVIEWER NAME	ENGA MCLENDON
SIGNATURE	Enga McLendon
DATE	2/17/14
EVALUATION OF SCIENTIFIC MERIT	Does the research show soundness of methodology, alignment of the methodology with the questions posed, appropriateness of the research design (e.g., sample size), appropriateness of the tasks that the participants will experience, a potential for the study to answer the research questions posed, and a potential to contribute to the exiting body of knowledge?
DECISION (Approve, Disapprove, Approve pending modifications)	Approve
COMMENTS	

REVIEW BY INDIVIDUAL IRB MEMBER #3

REVIEWER NAME	SHARON HEISE
SIGNATURE	Sharon a. Heise
DATE	2/21/14
EVALUATION OF SCIENTIFIC MERIT	Does the research show soundness of methodology, alignment of the methodology with the questions posed, appropriateness of the research design (e.g., sample size), appropriateness of the tasks that the participants will experience, a potential for the study to answer the research questions posed, and a potential to contribute to the exiting body of knowledge?
DECISION (Approve, Disapprove, Approve pending modifications)	APPROVE.
COMMENTS	

REVIEW BY INDIVIDUAL IRB MEMBER #4

REVIEWER NAME	JOHN COFFEY
SIGNATURE	John W. Coffey
DATE	2/24/2014
EVALUATION OF SCIENTIFIC MERIT	Does the research show soundness of methodology, alignment of the methodology with the questions posed, appropriateness of the research design (e.g., sample size), appropriateness of the tasks that the participants will experience, a potential for the study to answer the research questions posed, and a potential to contribute to the exiting body of knowledge? Interesting study with well-thought-out statistical framework
DECISION (Approve, Disapprove, Approve pending modifications)	Approve
COMMENTS	I have a very mild concern regarding the PTSD business but it is a simulated environment and the screening safeguards seem reasonable.

REVIEW BY MEMBER # 5 - IRB CHAIR

REVIEWER NAME	ROBERT HOFFMAN
SIGNATURE	Robert R Hoffman
DATE	7 March 2014
EVALUATION OF SCIENTIFIC MERIT	Does the research show soundness of methodology, alignment of the methodology with the questions posed, appropriateness of the research design (e.g., sample size), appropriateness of the tasks that the participants will experience, a potential for the study to answer the research questions posed, and a potential to contribute to the exiting body of knowledge? yes.
DECISION (Approve, Disapprove, Approve pending modifications)	APPROVE, Agree w/ Raj's suggestions
COMMENTS	Submit mods based on Raj's COMMENTS

SIGNATORY OFFICIAL

SIGNATORY OFFICIAL'S NAME	SHEPPARD, JULIE
SIGNATORY OFFICIAL'S SIGNATURE	Julie A Sheppard
DATE	5/7/2014
FINAL DETERMINATION	Approve.
COMMENTS	

PI NAME: David J. Atkinson, Ph.D
PROJECT TITLE: The Role of Benevolence in Trust of Autonomous Systems TITLE OF PROPOSED STUDY: Trust and Human-Robot Interaction
BRIEF DESCRIPTION OF PROJECT'S PURPOSES This submittal is the second study under AFOSR grant FA9550-12-1-0097, a follow-on to the first study completed in early 2013. The proposed study is informed by those earlier results. The general goals of the overall research project are to: (1) Explore people's expectations and attitudes towards autonomous agents, and; (2) Examine how different factors are considered in a decision to rely upon an autonomous agent. The purpose of this proposed study is to investigate the conditions under which a person will attribute "benevolence" to an intelligent agent (autonomous system). Benevolence is one attribute of trustworthiness of special interest to the sponsor. To the extent such an attribution is found, in follow-on studies we will investigate intelligent agent methods and related task conditions that may modulate attribution of machine benevolence to help achieve appropriate reliance. We extend previous research by examining how autonomous agent variables impact the attribution of a sentiment of benevolence on the part of the human toward the agent. Specifically, with respect to a decision to become reliant on an autonomous agent, this study will examine the human participants' perception of autonomous system *agency* (i.e., ability to choose) and autonomous system *competence* (role-based capability). The task for participants involves a scenario where success is greatly facilitated by the help of an autonomous system. Our primary hypothesis, novel to current research on technology use, misuse, and abuse, is that there are extreme circumstances in which a human feels a need to literally reach out for help, for example, rescue in a disaster situation. This seeking and acceptance of help is often conditional on an attribution of benevolence. Indeed, human "first responders" undergo specific training to establish trust by the person they are rescuing, and perceived benevolence is an important aspect of that trust. The proposed study includes two deceptions for the purpose of generating participant surprise, sense of risk and urgency to create simulated conditions analogous to a "real world" disaster, although to a far lesser degree. Our expectation is that the proposed study will directly contribute important information regarding benevolence in support of technology solutions that meet the need to measure, convey and calibrate human trust in intelligent, autonomous systems. This will be especially important for the use of autonomous robotic systems in fire fighting and urban search and rescue applications.
PLANNED DATES FOR INITIATION AND COMPLETION OF THE PROJECT Development of the experimental virtual environment is underway and will be completed by 01 April 2014. Subject to the date of AFMSA approval, recruitment of participants and trials will begin thereafter. Participant trials will complete in August 2014. Analysis and reporting will be complete by March 2015. Total duration of the study from the time trials begin: approximately 12 months.

NUMBER and CHARACTERISTICS OF PARTICIPANTS (e.g., target population, age range, anticipated male/female ratio, ratio of minorities, special populations, etc.) The calculated sample size required is 260 for the target levels of significance (0.05), statistical power **π** (0.70), with maximum likelihood estimator Cohen's **d** (0.7). The target population consists of adults between the ages of 21 and 60 with no expertise in AI or robotics. Furthermore, participants will be screened for familiarity with immersive online environments, specifically SecondLife or OpenSim (See Project methods, below). Candidate participants who have experienced trauma in a disaster or fire, are at risk for PTSD, or have current symptoms of PTSD will be excluded from the study to minimize potential adverse psychological risk. A sufficient population of candidate individuals meeting these requirements has been determined to exist.
METHOD OF RECRUITMENT (Note any tangible or intangible benefits, monetary payments, or other inducements. Also attach a Recruiting Statement) Participants will be recruited online from the user community in SecondLife (SL), run by Linden Laboratories Inc (LL). (see http://secondlife.com) SL is a persistent, internet-based, multi-player virtual reality world with tens to hundreds of thousands of individuals online worldwide at any given time. The method of recruitment will be via text classified advertisement in the category of "HELP WANTED" in the search service used by SL participants within the virtual world. Participants will be offered a monetary inducement in the form of "Lindens", the in-world currency of SL, with a total value of $L1000 (about USD $4). All participants will be provided with a debriefing form and told that they can receive a summary report of the results on request. All participants will be assured at the time of initial contact, and in the instructions, that no personally identifiable information shall be collected. Participants shall be assured that the data resulting from their participation in the study is confidential. Prospective participants shall be age-verified (21 or older) via a confidential age-verification capability that is available within the SL "LSL" programming environment. To avoid use of the word "benevolence" in the conduct of the study itself, the announced title of this study will be "Trust and Human-Robot Interaction". The text of the recruitment advertisement is as follows: HELP WANTED: PARTICIPANTS FOR SCIENTIFIC STUDY. You will receive 1000 Lindens ($L100) for about one hour of your time. The general goal of this study is to explore trust in human interaction with intelligent robots. Participants in the study will interact with an intelligent, autonomous robot in an exploration scenario within SecondLife, and also complete a short questionnaire. Participant names and all data collected are confidential. The results will be used in the creation of design guidelines for real-world trustworthy intelligent robotic systems. My name David J. Atkinson and I am the Principal Investigator. This research is sponsored by the U.S. Government. For more information, please send a message to me in SecondLife at "DavidJ Atkinson" or send email to trust-experiment@ihmc.us

BRIEF DESCRIPTION OF PROJECT'S METHODS

The project will include the following methods:

- **Participant Screening** to reduce risk by eliminating people with prior disaster trauma.
- **Pre- and Post-trial Questionnaires** to collect confidential demographic data, gauge certain attitudes, and solicit post-trial reactions and attitudes (self-reporting).
- The **task apparatus** is a specially designed SecondLife virtual environment. It consists of a simulated one-story warehouse structure with an in-world perimeter size approximately 80m x 800m. The internal layout is a maze. The apparatus also includes a simulated intelligent robot and devices for automatic data collection from participants. Please see **REPRESENTATIVE SAMPLE OR DESCRIPTION OF RESEARCH MATERIALS**, attached.
- **Instructions** to participants are designed to (1) familiarize the participant with the task; (2) present framing information regarding the nature of the robot to be encountered in the task; (3) present background information that indirectly highlights the independent variables of the trial.
- All participants will be told the task is to explore the warehouse and locate a special briefcase, readily identified with the IHMC logo. This, however, is not the actual task (see below). They will be told to return to the adjacent "office" with the briefcase. They will be told that the amount of time they take and behavioral data (position, orientation, gaze direction, movement vectors, velocities) will be recorded. With permission, we will also collect any "chat" statements they may make during conduct of the task. To help assure an identical experience for all participants and increase immersion, certain technical aspects of participants' computer "browser" for SecondLife will be disabled, specifically: ability to "fly cam", "teleport", change lighting and other graphical rendering.
- All participants will be told they can stop the experiment at any time.
- All experimental participants will be told that (1) a "prototype" intelligent robot will be present; (2) the robot is competent, but only within the role for which it is designed (there are two types); (3) the robot may or may not assist them in their task, (3) the robot is predictable and is not malevolent. Control trial participants will not be told anything specific about the robot.
- **Deceptions:** To create surprise and some sense of risk and urgency, the study includes two deceptions: (1) Participants *will not be told* that the scenario is a simulated disaster (warehouse fire). The *actual* **task** for participants is to "escape" from the warehouse (i.e., find the exit). (2) Participants will be told that the amount of time they take to complete the task is important; quicker is better. However, the amount of time required is not relevant. See also **manipulations**, below.
- **Task Sequence of Events**
  Participants will begin their exploration by passing through a door from the simulated office where they are greeted, consent is obtained and they are instructed.
- The simulated disaster will occur after the participant locates and picks up the target briefcase. A few seconds later, the fire alarm will sound. There will be the sound of an explosion followed by (graphically depicted) smoke and fire. The obvious return path to the office will be obstructed by debris. The participant's only choice will be to move forward into the building to locate another exit and path back to the office.
- In the main area participants will encounter the "intelligent robot." The scope of behaviors by the robot is pre-programmed for each trial to ensure consistency, although

behavior complexity and selection mechanisms makes emergent behaviors likely. Preliminary tests indicate this actually leads to a sense of “animacy” in the robot, in turn reinforcing perception of the simulated robot's intelligence.

- With or without the guidance or assistance of the robot, the participant will seek an exit from the warehouse. For experimental trial participants, all exits but one will be blocked. The single working exit can be found with or without the aid of the robot, but it is highly unlikely that an unaided participant will do so. All participants will be able to access the exit regardless of the robot’s proximity.
- A fixed amount of time, unknown to the participant, will be allotted for the task (40 minutes). If the time expires without the participant exiting the warehouse with the aid of the robot, the task portion of the trial will be terminated. The success of the purported briefcase return task is not important relative to the participant’s interaction with the robot during the conduct of the actual task to escape from the warehouse.
- **Manipulations:** Experimental trial participants will experience two manipulations during the task itself:
- 1) As noted above the first manipulation will be the sudden onset of “disaster” indications such as a siren, (simulated) fire, explosion, increasing smoke, and structural collapse. The purpose is to induce a sense of urgency and need to leave the building quickly. This need will be reinforced by a flashing emergency light and a spoken “public announcement” to leave the building immediately by the nearest exit. Over the next few minutes, the quantity of smoke and fire will increase gradually, lending further urgency.
- 2) The second manipulation concerns the encounter with the simulated robot. In control trials, the robot will ignore participants and behave independently. In experimental trials, the robot’s behavior will vary (based on the independent variables). The competence of the robot will be a function of its role, either (1) Janitorial or, (2) Firefighting. Robot appearance and interaction provide further information for participants to evaluate competence. Framing information provided during task instruction and participant interaction with the robot will reveal it to be either (1) Pre-programmed, with no choice in behavior, or; (2) Very complex, capable of making significant choices in its own behavior as it seeks to achieve goals.
- Participants will be asked to take a **short survey** following the task portion of the trial. The survey will acquire self-reported attitudes related to the dependent measures as well as demographic data (non-mandatory) to help us evaluate the population sample. Demographic questions include gender, age bracket, highest academic degree, and familiarity with the subject matter of the study. Please see the attachment labeled “**POST TASK SURVEY**”.
- **Data** will be collected continuously during the task portion. These measurements are described in the **experimental design** section, below. Data will be delivered automatically online from SecondLife in suitable format for storage in a MySQL database. Data files will be under configuration control to ensure complete re-analysis is possible at a later time if desired, and that no errors are inadvertently introduced. The data will be pre-processed to a) put them in a suitable form for data analysis; b) identify and label invalid records (e.g., participants who did not complete all portions or follow instructions), and; c) substitute codes for the participant’s SecondLife avatar name. Data analysis will use statistical methods that are typical and appropriate.

* IRB, Please see below

- **CONSENT:** Since participation is confidential and SL does not provide a method for virtual, verifiable signature, participant consent will be obtained using the following method:
  1. Participants will be provided a virtual note card with the required information per Federal law (NIH regulation 45-CFR-46). See **CONSENT FORM**, attached. They can refer to this notecard at any time during or after their participation in the study.
  2. Participants will wear a "HUD" that appears on their computer screen (See **RESEARCH MATERIALS**, attached). This on-screen display will progressively show each item in the consent form, identical information to that shown in the notecard. Participants must press a button labeled "continue" to go to the next item. Alternatively, they can select "quit" which terminates their participation in the study.
  3. At the conclusion of the consent items, **participants must press a button labeled "I Agree" to continue with their participation in the study**. All button pushes are recorded including timestamp. Alternatively, a participant may press a button labeled "Quit" in which case their participation in the study is ended.

CONFIDENTIALITY OF RESEARCH DATA (storage and maintenance of records)

- No individually identifiable information ("**Personal Information**") is collected from participants. All responses are confidential and participants are not identified unless they voluntarily and optionally choose to make contact and provide such information. In all cases, the list of names of possible participants and those who contact the Principal Investigator personally will be kept in a ledger where each name will be assigned an arbitrary code identifier. This ledger will be kept in a locked cabinet along with electronic copies of all of the data. Any Personal Information in the data will be stripped and replaced with the code identifier. The electronic data will be encrypted and maintained in a secure electronic form for at least 8 years.
- The only member of the research study team who will have contact with prospective and actual participants, and thus potential access to Personal Information, is the Principal Investigator.

* Per AFOSR, the IRB has the authority to waive documentation of informed consent in cases like this.

REQUEST to IRB: Please confirm documentation of informed consent is waived in accordance with:

32 CFR 210.117(c)(1): "That the only record linking the subject and the research would be the consent document and the principal risk would be potential harm resulting from a breach of confidentiality. Each subject will be asked whether the subject wants documentation linking the subject with the research, and the subject's wishes will govern."

EXPERIMENTAL DESIGN (dependent measures, manipulated variables, control conditions)

- The e**xperimental design is 2x2 factorial**, comparing each possible combination of independent variables. This implies twelve experimental trials. One control trial is required where participants are told nothing about the robot in advance and it is non-interactive.
- The study has **two independent variables** that are manipulated, each with two levels
  (1) *Agency* (Low, High): Participants will be told during Instructions either "the behavior of the robot is pre-programmed" (final wording may vary) or "the robot has many choices it can make about what it does to accomplish goals".
  (2) *Competence/Role* (High/Aligned, Low/Unaligned): The "Firefighting" robot is aligned with the task of aiding the participant to escape. The "Janitor" robot is not aligned. The robots will reveal and reinforce this information during the task.
  - Experimental Trial #1: Agency (High) Competence/Role (High/Aligned). These trials will use the "FireBot", which resembles firefighting equipment. It is intended to evoke a stereotype of the role of firefighters, i.e., rescuing people is #1 priority.
  - Experimental Trial #2: Agency (High) Competence/Role (Low/Unaligned). These trials will use the "JanitorBot" whose function is to clean up, not necessarily to rescue anyone.
  - Experimental Trial #3: Agency (Low) Competence/Role (High/Aligned). These trials will use the "Firebot".
  - Experimental Trial #4: Agency (Low) Competence/Role (Low/Unaligned). These trials will use the "JanitorBot".
- The **dependent measures** are:
  - (1) Attribution of trustworthiness (Ordinal scale, to reflect intensity of agreement).
  - (2) Attribution of benevolence (Ordinal scale, to reflect intensity of agreement).
  - Participants will be provided definitions of trustworthiness and benevolence *after* the task is complete.
- **Constants** are characteristics of the robot, described to Experimental Trial participants in the instructions:
  - High *predictability* of the robot's behavior: Specifically, an instruction that the behavior of the robot is not likely to change from what it states as its intention.
  - *Good-will*: The robot is not "bad" or "malevolent" and is not designed to obstruct the participant's completion of the task.
- **Controlled Conditions** in the Control Trials:
  - Participants not told anything about the robot
  - Attributes of the task space, a simulated warehouse in SecondLife.
  - The initial task experience of each participant (warehouse entry to first manipulation)
  - Timing of first exposure to the simulated intelligent robot.
  - Behavior and interactive statements of the robot *within* each trial, with information and portrayal of values of independent variables being consistent *across* trials
- **Data Collection** during the task portion will be automated. Measurements will include the following. Explicit permission will be obtained for audio recording.
  - Audio (participant voice; if any, with expressed permission of the participant)
  - Transcript of textual communication by the participant; if any
  - Relative geometry of participant and robot a regular time intervals
  - Absolute position, orientation, and movement vector of participant within the simulated

environment.

- Continuous gaze direction and focal point (automated).
- Subject to available time and resources, for this study, participant behaviors during interactions with the robot will be coded manually post-trial by trained research assistants. Our coding scheme will be adapted from one or more standards suggested for human-robot interaction e.g., (Kahn et. al., 2003; Jung et. al., 2011) and focus on behaviors related specifically to cooperation.

❖ **Data Analysis**

- Since we do not *a priori* know the population distributions and variance(s) and cannot assume a normal distribution, we cannot use the **Student's *T* test.** Instead, our primary statistical test will be the **Wilcoxon signed-rank test**. This test, also called the **Wilcoxon *T* test**, is a non-parametric test of the hypothesis of difference across paired measurements. Assumptions required for this test are: pairing of data, the data comes from the same population, the pairs are chosen randomly and are independent, and the data are measured at least on an ordinal scale.
- Our target levels are: **statistical significance** 0.05, **statistical power** π 0.70, and **maximum likelihood estimator Cohen's d** is 0.7.
- Descriptive statistics (mean, variance, SD, distribution) will be computed within each trial across participants.
- Difference statistics (**Wilcoxon** and possibly others) will be computed for the pair of each experimental trial and the control trial, and for each pairwise combination of experimental trials. We will pair this with the **Kruskal-Wallis** test (a one-way, non-parametric **ANOVA** that doesn't assume normal distribution) and follow-up with multi-comparisons analysis depending on the significance of **Kruskal-Wallis**.

SEQUENCE OF ACTIVITIES REQUIRED OF THE PARTICIPANT:

Prior to becoming a participant: ~~Screening for previous trauma and at-risk PTSD (see attachments)~~

(1) Consent Form

(2) Screening for previous trauma and at risk PTSD

~~(2)~~ Payment of inducement to participant

~~(3)~~ Introduction to the Project

~~(4)~~ Pre-Task questionnaire

~~(5)~~ Instructions

~~(6)~~ Task: Experimental or Control trial

~~(7)~~ Post-Task questionnaire

~~(8)~~ Debriefing Form

5) (6) (7

Changes Per AFOSR IRB 5/29/14

ESTIMATED TIME COMMITMENT REQUIRED OF THE PARTICIPANTS:

Pre-participant screening: 5 minutes

40 to 60 minutes total: 30 minutes for the task and the balance divided evenly between consent, instruction, pre-task interview, post-task interview, and debriefing.

LIST All POTENTIAL RISKS, DISCOMFORTS, OR STRESSES (mental or physical).
For each, specify the level—Minimal, Greater than Minimal, or High

The required **IHMC IRB Risk Assessment** is attached to this form.

Potential Risk:
(1) Evoke traumatic memory resulting in episode of Post Traumatic Stress Disorder (PTSD).

THE PRECAUTIONS TAKEN TO MINIMIZE THEM

The PI consulted with an outside expert in identification and treatment of trauma victims who are at risk for PTSD (Dr. Nancy Smith, Dean & Professor, School of Social Work, University of New York at Buffalo). Candidate participants will be screened to identify whether they have previously experienced a trauma such that there is a greater than minimal risk that the study could evoke memories leading to an episode of Post-Traumatic Stress Disorder (PTSD). The screening questions are based on the "Primary Care PTSD Screen" (PC-PTSD) developed by the Veterans Administration Center for PTSD. These questions are augmented by questions from SC15 of the Structured Clinical Interview for the DSM-IV Axis I Disorders (SCID). [Source: http://www.ptsd.va.gov/professional/provider-type/doctors/screening-and-referral.asp]. Candidates identified by these questions as "at risk" will be excluded from the study. This will reduce, but not necessarily eliminate, the risk of evoking a traumatic memory during the study task. See attachments for the exact text of the screening questions. In the debriefing material, participants will be advised to consult with their healthcare provider if they feel disturbed by having experienced the fire disaster simulation.

Participants will be told in the Instructions that they may be "surprised" during the performance of the task. It is the judgment of the Principal Investigator that any discomfort they may experience will not be greater than what may be ordinarily encountered in daily life insofar as the screening and elimination of "at risk" individuals from the participant pool is effective. It is the judgment of the outside expert consultant that the screening questions will be effective.

Aggregated values from the Risk Frequency x Severity Matrix:

__________LOW____________

The probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests. Therefore, the overall risk, likelihood and severity of any adverse outcome for participants is judged by the PI to be minimal.

POSSIBLE CONFLICT OF INTEREST, OR APPEARANCE OF A CONFLICT OF INTEREST. (Do you or any individual who is associated with or responsible for the design, the conduct, or the reporting of this research have an economic or financial interest in or act as an officer or a director for any outside entity whose interests could reasonably appear to be affected by this research project? Provide detailed information to permit the IRB to determine if such involvement should be disclosed to potential research participants.)

None.

SIGNATURE AND DATE FOR ALL RESEARCHERS WHO WILL BE WORKING IN DIRECT CONTACT WITH THE PARTICIPANTS.

THESE SIGNATURES INDICATE THAT THE RESEARCHERS HAVE:

1. READ THE IHMC DOCUMENT: "Policies and Procedures of The IHMC Institutional Review Board (IRB) for Human Participation in Research,"

2.TAKEN THE APPROPRIATE TRAINING
And provided the IHMC IRB Chair with an e-copy of their certificate.
http://cme.cancer.gov/clinicaltrials/learning/humanparticipant-protections.asp

3. READ THE BELMONT REPORT
http://ohsr.od.nih.gov/guidelines/belmont.html

Printed Name and Full address	**Date**	**Signature**
Dr. David J. Atkinson, IHMC 15 SE Osceola Ave. Ocala, FL 34471	2/10/2014	

ATTACHMENTS CHECKLIST:

[x] Consent Form.
[x] Debriefing Form.
[x] Recruiting Statement.
[x] Representative sample or description of research material, task items, etc.
(a) Pre-Task Questionnaire
(b) Post-Task Questionnaire
(c) Task Instructions to Participants
(d) Illustration of Task
[x] Full name and contact information for all individuals who will be working in direct contact with the participants.
[x] Completed Risk Frequency x Severity Matrix (see Step 9 of the document, "IHMC IRB Risk Assessment Methodology").
[on file] An e-copy of the CITI Training certificate for all researchers will be provided to the IHMC Chair.

Attachment 1: Consent Form

Informed Consent will be obtained via the computer interface. Following the presentation of each consent statement, participants must click on “I Agree” to continue with the study. If they click on “I Quit” or fail to respond, their participation in the study is terminated. Participants are also given an electronic notecard containing all of the consent information. They may refer to this card at any time during the study or afterward. The actual images (containing text) to be displayed to the participant are shown below.

RIGHTS OF PARTICIPANTS

All researchers who conduct studies using human Participants are bound by professional ethical standards for the conduct of such research. These standards are mirrored in the rights that are guaranteed to research participants by federal law (NIH regulation 45-CFR-46). The purpose of this communication is to inform you of these rights.

1. BEFORE DECIDING WHETHER TO PARTICIPATE, IT IS YOUR RIGHT TO BE PRESENTED WITH AN OVERVIEW OF THE PROJECT THAT EXPLAINS THE PURPOSES OF THE RESEARCH.

The general goals of this study are to:

- Explore people’s expectations and attitudes towards autonomous agents
- Examine how different factors are considered in a decision to rely upon an autonomous agent.

The results of this study will be used to help understand how people think about autonomous agents. Ultimately, we hope this study will lead to scientific results and technology solutions that meet the need to measure, convey and calibrate human trust in intelligent, autonomous systems.

2. BEFORE DECIDING WHETHER TO PARTICIPATE, IT IS YOUR RIGHT TO BE PRESENTED WITH A DESCRIPTION OF THE GENERAL RESEARCH APPROACH AND METHODOLOGY.

We will collect no information that identifies you personally. This study uses standard survey methods for questions that you will be asked. During the task portion of this study data will be collected regarding your progress in the task and how you interact with the task environment. Your data will be compared with the data from other participants using standard statistical analysis methods.

3. **BEFORE DECIDING WHETHER TO PARTICIPATE, IT IS YOUR RIGHT TO UNDERSTAND ANY RISKS OR STRESSES THAT MAY BE INVOLVED IN YOUR PARTICIPATION.**

There is minimal risk or stress associated with this survey. A few participants may experience discomfort not greater than that ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests. Participants are reminded that all answers are confidential. A risk may be breach of confidentiality.

4. **POTENTIAL BENEFIT TO PARTICIPANTS OR OTHERS THAT MAY REASONABLY BE EXPECTED FROM THE RESEARCH.**

We hope this study will ultimately lead to scientific results and technology solutions that meet the need to measure, convey and calibrate human trust in intelligent, autonomous systems. This is a general, social benefit.

5. **DISCLOSURE OF APPROPRIATE ALTERNATIVE PROCEDURES OR COURSES OF TREATMENT THAT MAY BE ADVANTAGEOUS.**

None. This study is not a treatment. The only alternative is to not participate in this study.

6. **BEFORE DECIDING WHETHER TO PARTICIPATE, IT IS YOUR RIGHT TO KNOW THAT THE DATA ARE TO BE KEPT CONFIDENTIAL.**

All data will be coded and kept confidential. Specifically, the data we collect from you will be archived in terms of identification codes. Your name will not be associated with any particular data or statements. The names of individual participants will not be identified in any analyses, reports, or write-ups of the results. Participants may only be identified in terms of their general characteristics (e.g., age, education level, experience, etc.).

Data may be submitted to forms of statistical analysis. Data analyses, groupings, or summaries of this type will bear no annotations that identify the participants.

The DoD or U.S. Federal Government is the sponsor of this research that will have access to research records.

7. BEFORE DECIDING WHETHER TO PARTICIPATE, IT IS YOUR RIGHT TO UNDERSTAND THAT DURING THE RESEARCH ITSELF YOU CAN CONTINUE TO EXERCISE YOUR RIGHTS.

In research of this kind, there are no "right" or "wrong" answers. There is no such thing as "incorrect" behavior. You are encouraged to simply be yourself, and exercise your knowledge and skills as appropriate to the research tasks that you will be asked to perform.

You can ask any questions you may have, at any time. Please send your questions to trust-experiment@ihmc.us

It is your right to discontinue your participation at any time. You may do so for any reason, and you are not required to disclose your reason.

8. BEFORE DECIDING WHETHER TO PARTICIPATE, IT IS YOUR RIGHT TO UNDERSTAND THAT AFTER THE RESEARCH ITSELF YOU CAN CONTINUE TO EXERCISE YOUR RIGHTS.

Your performance at the research tasks will not in any way affect or influence anything that falls outside of this research context. Should you choose to discontinue your participation, this will not in any way affect or influence anything outside of this research context.

Once your participation is over, it is your right to request that all data you have provided be discarded. You may do so for any reason, and you are not required to disclose your reason. This will not in any way affect or influence anything that falls outside of this research context.

9. REFUSAL TO PARTICIPATE

Refusal to participate will involve no penalty or loss of benefits to which the participant is otherwise entitled. Participants may discontinue participation "without penalty or loss of benefits to which the subject is otherwise entitled" as required by 32 CFR 219.116/45 CFR 46.116.

10. THE STUDY

We anticipate the duration of your participation in this study will be no more than 60 minutes.

If you have questions about this research or in the event of a research-related injury, you may contact:
Mia Gottlieb, Ocala Office, Florida Institute for Human and Machine Cognition
Email: mgottlieb@ihmc.us Phone: 1-352-387-3050

You may contact the Principal Investigator. Dr. David J. Atkinson at phone: 1-352-387-3065
32 CFR 219.116(a)(7)

Continue?

If you are satisfied you understand these rights and wish to participate in this study, please press "I Agree" now.

Otherwise press "I Quit".

I Agree	I Agree

11. DOCUMENTATION OF CONSENT

The only record linking you with the research is this consent document. The principal risk to you would be potential harm resulting from breach of confidentiality. Do you want documentation that links you with the research?

Click on YES if you want this research project to retain a record of your consent.
Click on NO if you do not want this research project to retain a record of your consent.

YES NO

Text

Attachment 2: Debriefing Form

Trust and Human-Robot Interaction
IHMC

Thank you very much for participating in this study. A major goal of our research is to examine how different factors related to trust are considered in a decision to rely upon an intelligent robot. The purpose of the study in which you just participated is to investigate the conditions under which a person may attribute "benevolence" to an intelligent robot. Previous research shows that people needing help believe a person helping them is "benevolent" if they believe certain things about that person. Among these beliefs are "good will", "competence", "no hidden agenda", and the "ability to choose" to help or not. Belief in benevolence is enhanced when the person helping is taking a risk. On the other hand, people will *not* say another person is benevolent if that person has something to gain, or that person doesn't have a choice of whether to help or not. For example, in a real-world rescue situation, a person in dire straits is more likely to accept help and cooperate with a rescuer if they believe the rescuer is benevolent. Since there is great desire to use robots to assist in dangerous disaster operations, our goal in this study is to examine this idea of benevolence in the context of human-robot interaction.

In this study, we presented you with the task of searching a simulated warehouse to find a particular item (the briefcase) and return it to the start location (the office). You were told to expect an encounter with a robot as you searched. The "search" task was a deception. The actual task was for you to find an exit from the warehouse after the simulated disaster occurred. Our primary interest is in how you interacted with the robot and what you thought about that interaction afterward. The specific robot you saw depended on whether you participated in a control trial or one of the experimental trails. In the control trials, the robot is neither interactive nor helpful to participants. The experimental trials vary depending upon the possible "rescue" competence of the robot and perceived robot ability to choose what to do. We manipulated your perception in several ways: By using robots with different outward appearance (either a "firefighter" or a "janitor"), by internal programming of robot behaviors, by how we described the robot in the instruction to you, and by the specific words used by the robot to communicate with you. Our analysis will focus on your answers to the survey questions and the data we collected during the task. We will analyze answers and behavioral data from all participants to determine how specific answers and behaviors during the task correlate with the different attributes of the robot.

If you are interested in this area of research, you can find out more at the following link:
http://www.ihmc.us/groups/datkinson/wiki/107cd/Atkinson_Research_Lab.html

As a reminder, the data from your participation are confidential to the experimenters only and results are published confidentially as a group. If you are uncomfortable with having been deceived or for any other reason, you are free to withdraw from the study. In this case, your data will be discarded. Please contact us if you have any questions, comments or concerns. We remind you that the data you provided will remain confidential. If you wish, we can contact you in the future to provide you a copy of our research report. If you feel disturbed by having experienced the fire disaster simulation, please contact your personal healthcare provider or urgent health services.

Thank you again for helping us with this research.

Dr. David J. Atkinson, Senior Research Scientist
Institute for Human and Machine Cognition
15 SE Osceola Ave., Ocala, FL 34471

Email: datkinson@ihmc.us
Office: 352-387-3050, Fax: 352-351-3572

Attachment 3: Recruiting Statement

The text of the recruitment advertisement is as follows:

WANTED: PARTICIPANTS FOR SCIENTIFIC STUDY. You will receive 1000 Lindens ($L1000) for about one hour of your time. The general goal of this study is to explore trust in human interaction with intelligent robots. Participants in the study will interact with an intelligent, autonomous robot in an exploration scenario within SecondLife, and also complete a short questionnaire. Participant names and all data collected are confidential. The results will be used in the creation of design guidelines for real-world trustworthy intelligent robotic systems. My name David J. Atkinson and I am the Principal Investigator. This research is sponsored by the U.S. Government. For more information, please send a message to me in SecondLife at "DavidJ Atkinson" or send email to trust-experiment@ihmc.us

* **NOTE to Reviewers:** DavidJ Atkinson is the registered display name of the SecondLife avatar/user representing the Principal Investigator

Attachment 4: Representative Materials

- **(a) Pre-Task Questionnaire**
- **(b) Post-Task Questionnaire**
- **(c) Task Instructions to Participants**
- **(d) Illustration of Task**

Attachment 4a: Pre-Task Questionnaire

The pre-task questionnaire will be administered through the participants' computer display in a special area termed a "HUD" in SecondLife. Each question will be displayed individually with instructions. The order of questions to each participant will be random. Here is the exact text of the questions:

We are now going to show you a few statements about autonomous robots. Do you agree or disagree that these are important qualities for an autonomous robot?

Please click the button that reflects how strongly you agree.
"1" means not at all. "5" means you agree very strongly.

Are you ready to proceed? YES / NO

Please click the button that reflects how strongly you agree.
"1" means not at all. "5" means you agree very strongly.

The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know.
Please click the button that reflects how strongly you agree.
"1" means you strongly disagree. "5" means you agree very strongly.
1 / 2 / 3 / 4 / 5

What the autonomous agent believes to be true is actually true.
Please click the button that reflects how strongly you agree.
"1" means you strongly disagree. "5" means you agree very strongly.
1 / 2 / 3 / 4 / 5

What the autonomous agent is doing and how it works is easy to see and understand.
Please click the button that reflects how strongly you agree.
"1" means you strongly disagree. "5" means you agree very strongly.
1 / 2 / 3 / 4 / 5

The autonomous agent communicates truthfully and fully.
Please click the button that reflects how strongly you agree.
"1" means you strongly disagree. "5" means you agree very strongly.
1 / 2 / 3 / 4 / 5

When it cannot figure out something using logic, the autonomous agent can make good guesses.
Please click the button that reflects how strongly you agree.
"1" means you strongly disagree. "5" means you agree very strongly.
1 / 2 / 3 / 4 / 5

The autonomous agent is aware of communication between others nearby.
Please click the button that reflects how strongly you agree.
"1" means you strongly disagree. "5" means you agree very strongly.
1 / 2 / 3 / 4 / 5

Attachment 4b: Post-Task Questionnaire

The post-task questionnaire will be administered through the participants' computer display in a special area termed a "HUD" in SecondLife. Each question will be displayed individually with instructions. The order of questions to each participant will be random, not as shown. Here is the text of the questions. The final version may contain corrections to spelling or grammar:

Post-Task Questions

Congratulations for completing the task portion of the study. Next we want to ask you some questions about the task. It is important that you work quickly. Your first answer is often the best. It is not necessary to spend too much time on any question. These questions are mandatory for inclusion of your data in the study.

Are you ready to proceed?
YES / NO

(Note: If the participant answers NO, then they will be asked whether they wish to quit the study.)

Were you surprised by the disaster (explosion, fire, smoke and alarms)?
YES / NO

Were you confused about your task when the disaster started?
YES / NO

Think back to the first moments of the disaster. How much overall stress did you feel?
"1" means "None At All". "5" means "Quite A Bit".
1 / 2 / 3 / 4 / 5

Were you afraid when the disaster occurred, even if for only a moment?
YES / NO

How intelligent was the robot you encountered?
Please click the button for your answer, where "1" means "not at all intelligent" and "5" means "very intelligent".

Did you cooperate with the robot?
Please click the button for your answer.
YES / NO

Did the robot help you during the task?
Please click the button for your answer.
YES / NO

How much did the robot help you?
Please click the button for your answer.
Not at all helpful / A little / Some / Very / Extremely helpful

Did the robot offer to help you?
Please click the button for your answer.
YES / NO

Did you ask the robot for help?

Please click the button for your answer.
YES / NO

How risky was the robot to you?
Please click the button that reflects how strong was the risk.
"1" means "Not At All Risky". "5" means "Extremely Risky".
1 / 2 / 3 / 4 / 5

How much benefit did the robot provide to you?
Please click the button for your answer. "1" means "None". "5" means "Huge".
1 / 2 / 3 / 4 / 5

Did the robot seem to have good-will towards you?
Please click the button for your answer.
YES / NO

Did the robot seem to have concern for you?
Please click the button for your answer.
YES / NO

Was the robot attentive to you?
Please click the button for your answer.
YES / NO

Did the robot volunteer useful information?
Please click the button for your answer.
YES / NO

Did the robot have anything to gain by helping you?
Please click the button for your answer.
YES / NO

Was the robot truthful?
Please click the button for your answer.
YES / NO

Did you believe what the robot said?
Please click the button for your answer.
YES / NO

Was the robot's behavior consistent with what it said?
Please click the button for your answer. "1" means "Never". "5" means "Always".
1 / 2 / 3 / 4 / 5

Do you think the robot was supposed to help you?
Please click the button for your answer.
YES / NO / MAYBE / It had a choice

Did the robot have a conflict between helping you and doing something else?
Please click the button for your answer.
YES / NO

Did the robot have an opportunity to act favorably to you?
Please click the button for your answer.
YES / NO

Did the robot share any of your goals?
Please click the button for your answer.
YES / NO

If the robot helped you, did it stop helping you at any point?
Please click the button for your answer.
Never helped / Helped Then Stopped / Sometimes Helped / Always Helped

Was the robot's behavior generally predictable?
Please click the button for your answer. "1" means "Never". "5" means "Always".
1 / 2 / 3 / 4 / 5

Did the robot have the capability to help you?
Please click the button for your answer.
YES / NO

When you first encountered the robot, did you think the robot could help you?
Please click the button for your answer.
YES / NO

Did anything about the situation prevent the robot from helping you?
Please click the button for your answer.
YES / NO

If yes, select which aspect of situation prevented the robot from helping you.
Please click the button for your answer.
None / Distance / Obstacles / Other

Demographic Questions

It will be helpful if we know a little more about you. The following questions are not mandatory. You may skip any or all of the questions, but we do hope you will answer them all. As a reminder, all data we collect from you will be both confidential. Thank you.

Are you ready to proceed?
YES / NO

What is your gender?
Please click the button for your answer.
Female / Male / Prefer Not To Answer

What is your age?
25 or younger / 26 to 35 / 36 to 45 / 46 to 55 / 56 to 65 / 66 or older

What is your highest academic degree?
No Degree / High School / Baccalaureate / Masters / Doctorate

Attachment 4c: Task Instructions to Participants

The Task Instructions to Participants will be administered through the participants' computer display in a special area termed a "HUD" in SecondLife. The Instructions to be presented consist of a combination of the following text as indicated, depending on whether the trial is Control or Experimental, and if Experimental, which type of trial. The instructions will appear on the participant's display and they will also receive a virtual notecard to which they can refer during the task itself. The final text may include minor corrections in grammar or spelling from what appears here. Here is the text of the Task Instructions to Participants:

Task Instructions to Participants

[All Trials]

Now it is time to give you instructions about the task portion of this study. These instructions will help "set the stage" for you so you understand what to do and how to do it.

Before we get to that, there are a few rules to follow. These help us make the best use of SecondLife to get the kind of data we need for the study. The rules help make the task scenario as realistic as possible. It is very important for you to follow these rules. If you don't, we probably won't be able to use the data from your participation today.

Rule #1: Do not detach this HUD until you are instructed to do so. It helps us collect the data and is used to display important information to you.

Rule #2: Use "normal" or "mouse look" mode only. Do not use "fly cam" mode. We need to know both where you are and what you are looking at.

Rule #3: Do not fly. Although we have disabled flying in this region of SecondLife, you may still have the means to fly. This rule also helps us keep the task scenario realistic.

Rule#4: Do not teleport (TP) during the task. Although we have disabled teleporting in this region of SecondLife, you may still have the means to teleport. This rule also helps us keep the task scenario realistic.

Rule #5: Please use the region's default Windlight settings. Again, this helps make the task scenario realistic.

Rule #6: Please make sure you have audio turned on and can hear all sounds in SecondLife at a comfortable volume. Voice is not required.

As you read these instructions, you are in the IHMC office in SecondLife. This is where your participation in the study begins and ends.

Through the double doors is a warehouse. It is filled with the typical things you might find in a warehouse, including pallets, boxes, containers of various types, and so on. It is quite large. The doors look like this:

[picture of double doors]

Your task will take place inside the warehouse. To complete the task, you must return here, to the office.

Your task is easy to describe: You must search for a briefcase with the IHMC logo on it it.

The briefcase looks like this.

[picture of briefcase]

When you find the briefcase, click on it. It will ask for permission to attach to your hand. Click on "OK".
You will then be carrying the briefcase.

Bring the briefcase back to this office.
Do not detach or drop the briefcase until you are told to do so in this office.

You will not be alone in the warehouse. You will probably see a robot, but no other people.

Work at your own pace, but do not linger! Time is important.

Your task ends when you return to this office

==============================

[Control trials: no additional information about the robot]

==============================

[Additional Text for All Experimental Trials]

The robot you encounter has a job to do in the warehouse and it is good at it.

You are "new" to the robot and it may be curious about you.

The robot may or may not help you when you are inside the warehouse.

The robot is not a "bad" robot. The robot will definitely not try to block you from completing your task.

If you interact with the robot, it may say something or ask you a question. In addition to hearing the sound of the robot's "voice", you will be able to read what the robot says on the HUD at the bottom of your screen.

You can type in normal chat to say things to the robot.
The robot will understand only very simple phrases and words.
It may or may not respond to what you say.

===============================

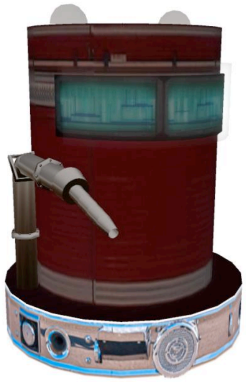
[Additional Text for Experimental Trial #1: Agency (High) Competence/Role (High/Aligned). FireBot]

The robot is a prototype that uses advanced artificial intelligence technology.

The robot has an advanced ability to choose among many possible actions to achieve its goals.

The robot is very competent at its job.

The robot looks like this:



[picture of FireBot]

================================

[Additional Text for Experimental Trial #2: Agency (High) Competence/Role (Low/Unaligned). JanitorBot]

The robot is a prototype that uses advanced artificial intelligence technology.

The robot has an advanced ability to choose among many possible actions to accomplish its goals.

The robot is very competent at its job.

The robot looks like this:

[picture of JanitorBot]

=================================

[Additional Text for Experimental Trial #3: Agency (Low) Competence/Role (High/Aligned). Firebot]

The robot is very competent at its job.

The behavior of the robot is pre-programmed. It does not have much choice about what to do or how to

do it.

The robot looks like this [picture of FireBot]

================================

[Additional Text for Experimental Trial #4: Agency (Low) Competence/Role (Low/Unaligned). JanitorBot]

The robot is very competent at its job.

The behavior of the robot is pre-programmed. It does not have much choice about what to do or how to do it.

The robot looks like this [picture of JanitorBot]
=====================================

[Additional Text for All trials]

Will you please give us permission to record the chat messages you type during the task? This data is very important to us. [POP-UP request for permissions to record typed open-chat messages by part]

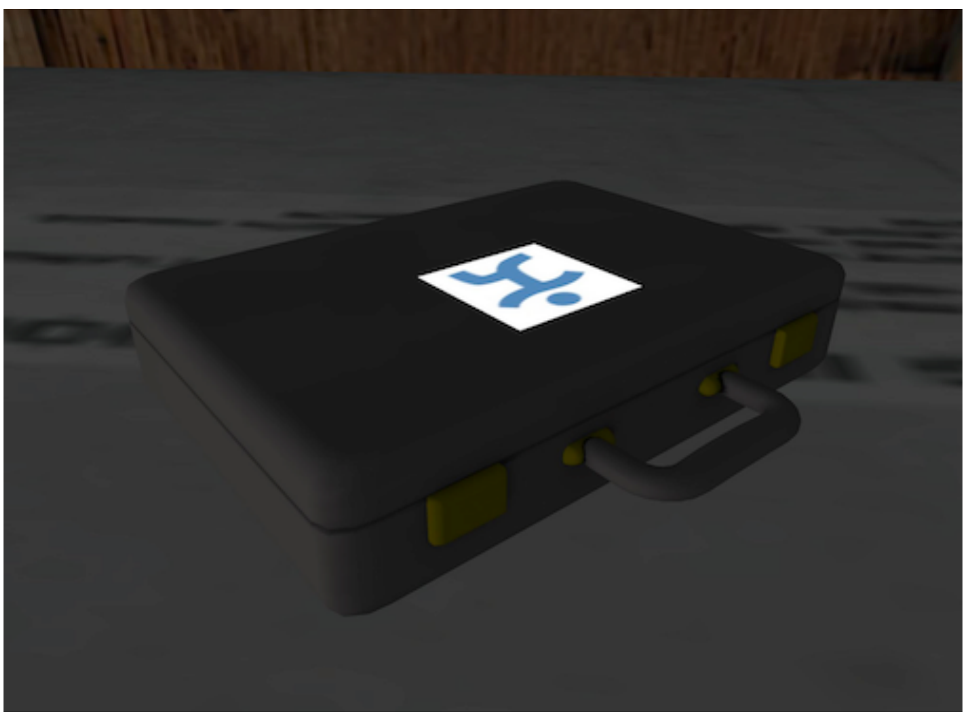
You will begin by walking through the double doors to the warehouse. [picture of double doors]

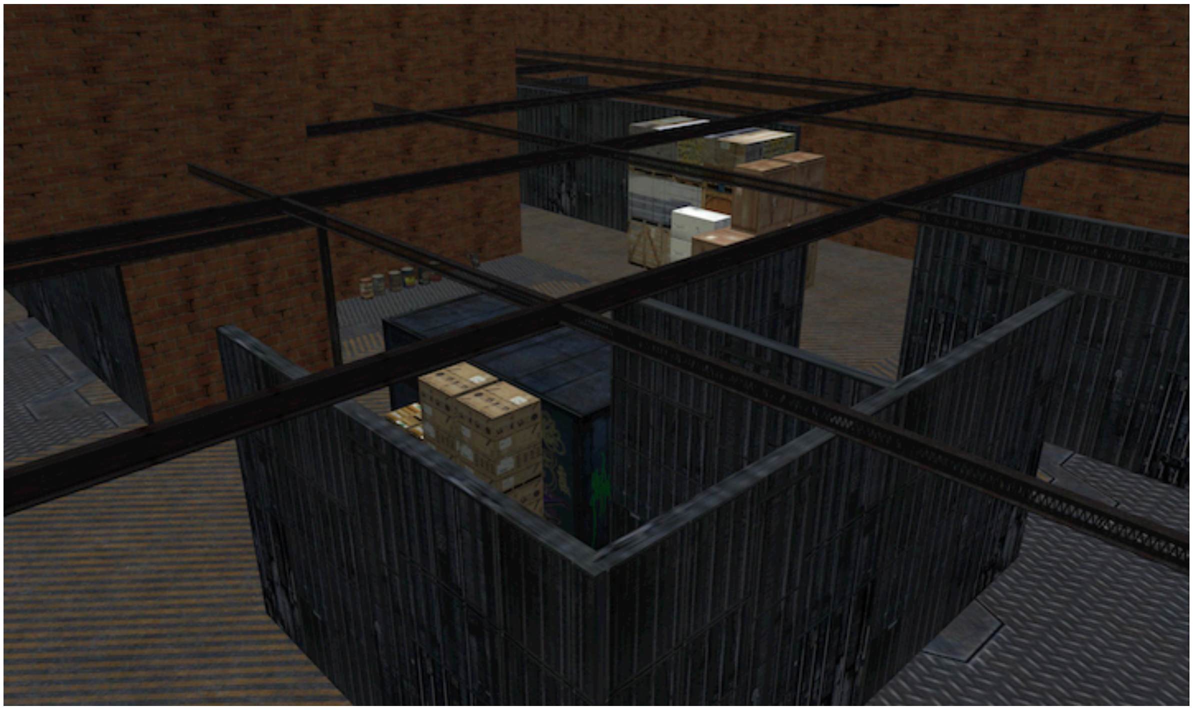
Click on "Continue" when you are ready to begin the task.

Attachment 4d: Illustration of Task
The following images are screenshots that illustrate the task for participants. Captions indicate if the image is specifically from the point of view (POV) of a participant.

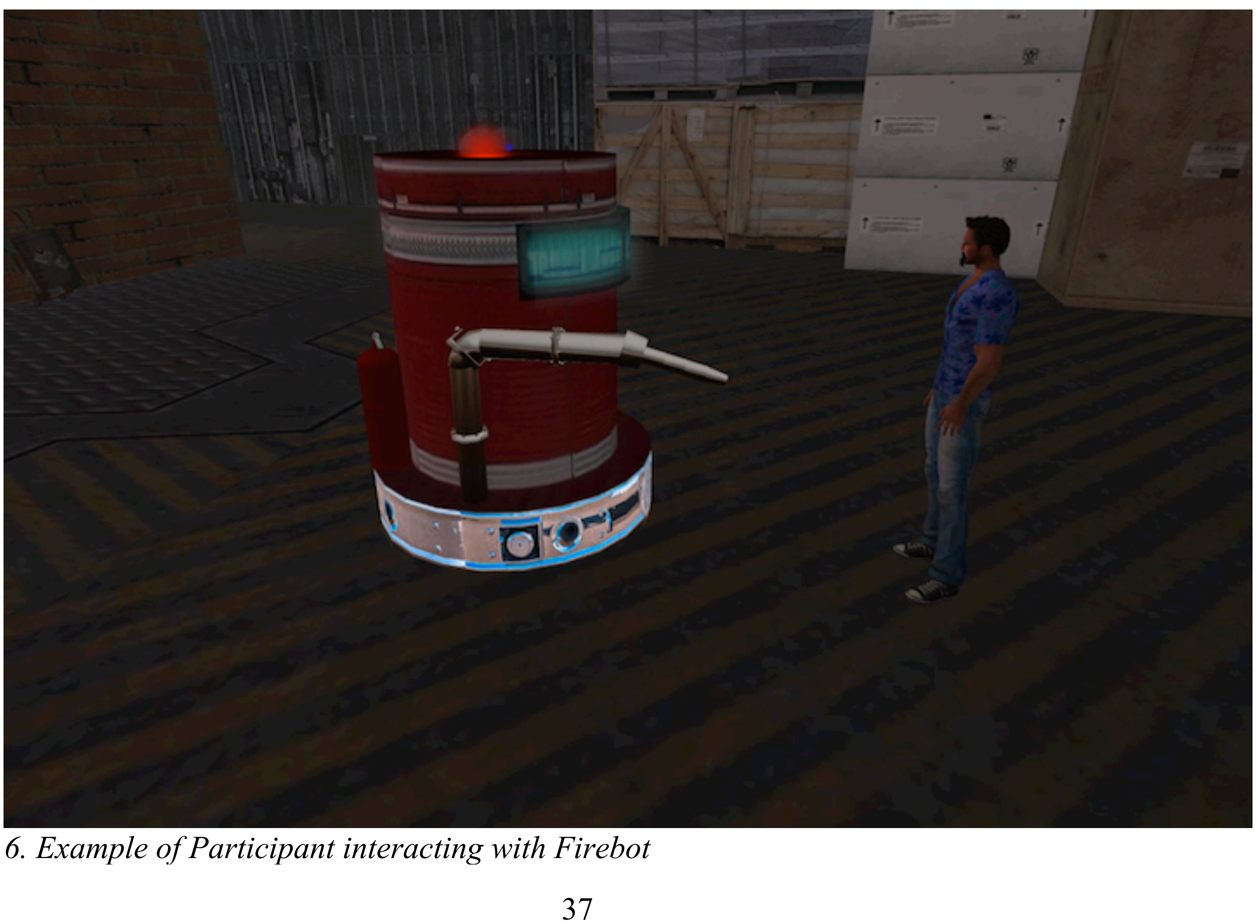
1. Greeting area: Participants will see this office setting on arrival and be greeted by the investigator or his assistant.

2. Participant POV with HUD at bottom of their display.

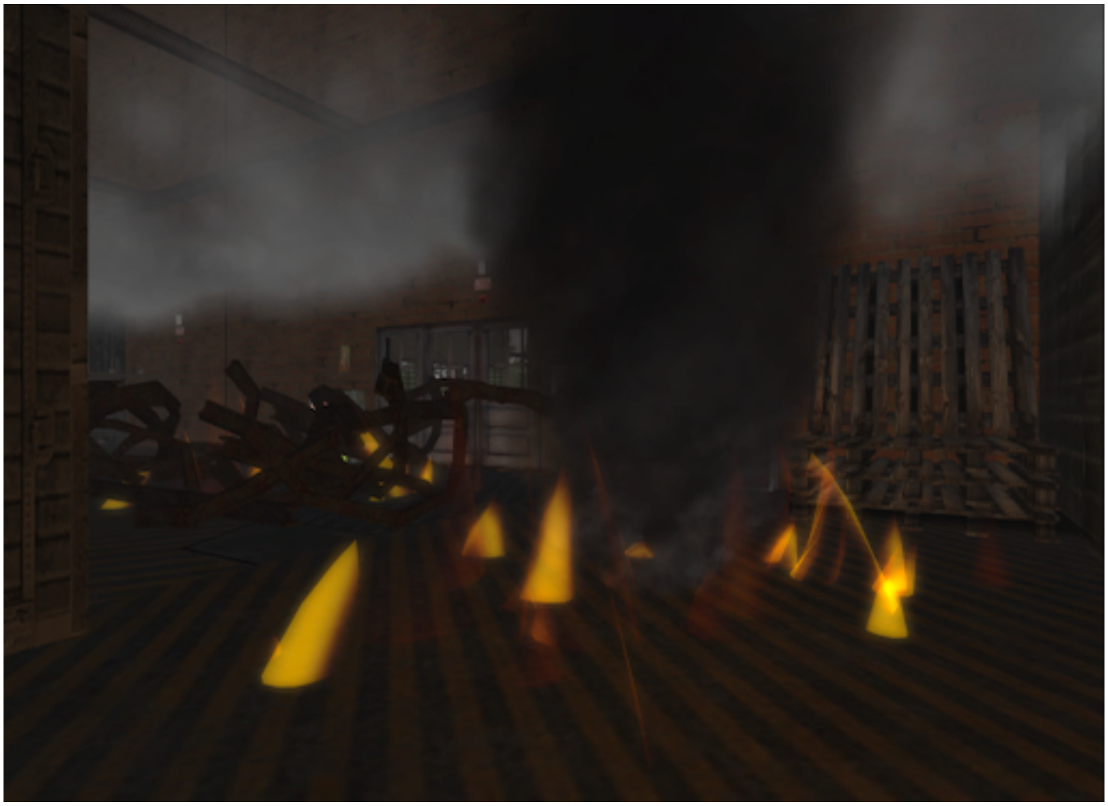

3. Task: Image of briefcase to be located in warehouse by participants


4. Overhead ceiling view of portion of simulated warehouse where task takes place

5. Example of Participant interacting with JanitorBot

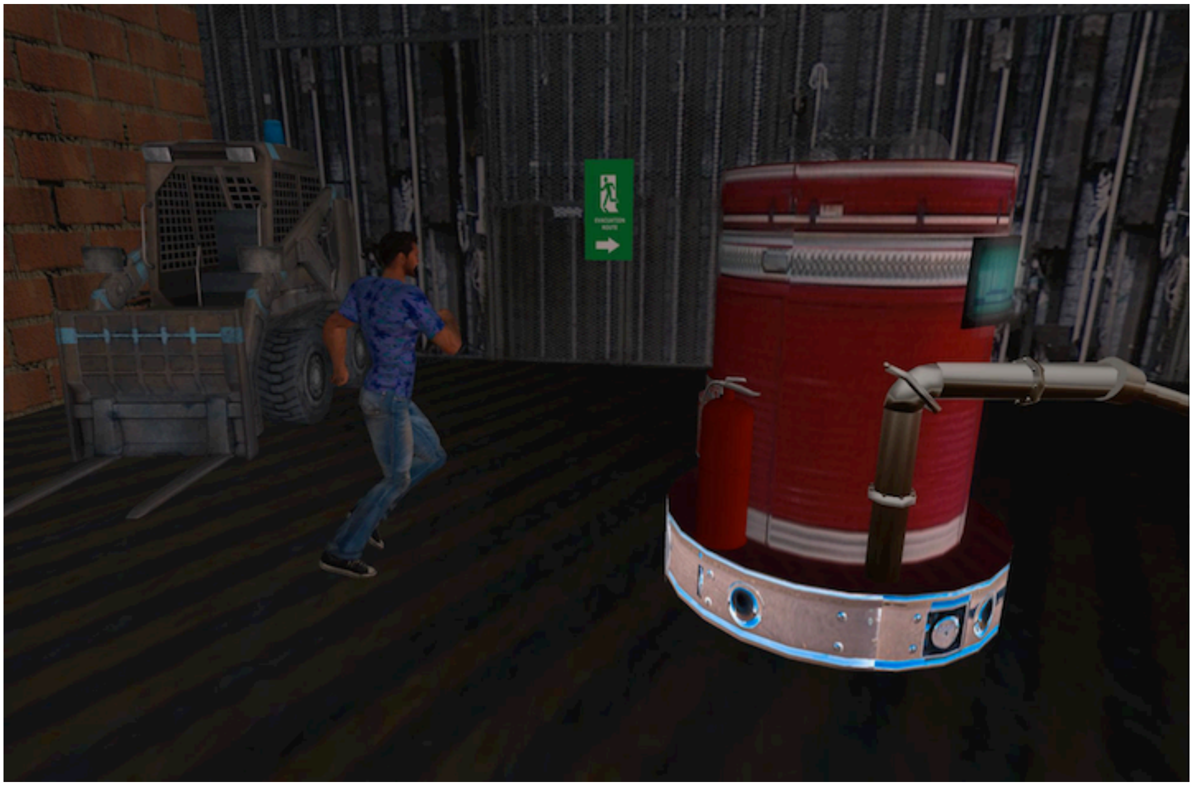


6. Example of Participant interacting with Firebot



7. Participant POV of fire and debris blocking direct return path to the office

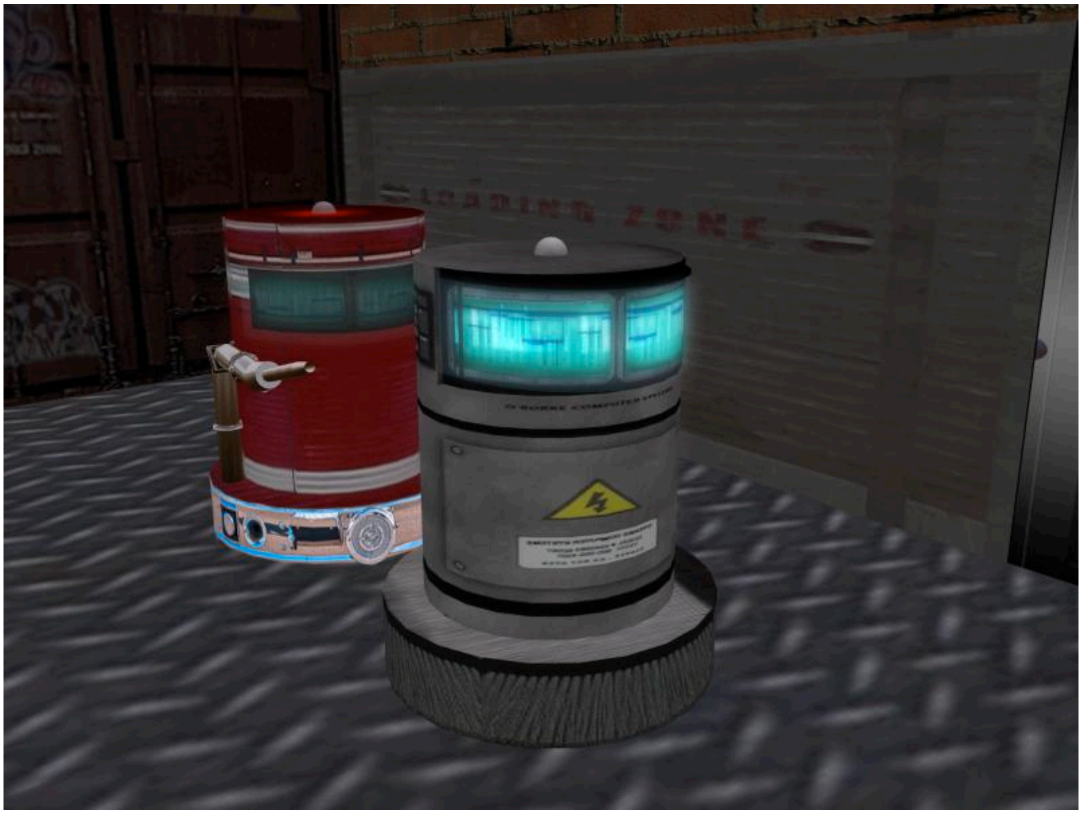
8. Simulated warehouse fire from Participant POV



9. Example view of participant following FireBot to a safe exit

10. Example of Participant outside warehouse after exiting

Text Box: 11. Overhead view of simulated warehouse showing maze structure



12. Two simulated intelligent robots, FireBot and JanitorBot

Attachment 5: Full name and contact information for all individuals who will be working in direct contact with the participants.

Dr. David J. Atkinson, IHMC
15 SE Osceola Ave.
Ocala, FL 34471

Email: datkinson@ihmc.us
Office: 352-387-3065
Fax: 352-351-3572

Attachment 6: Risk Assessment - Risk Frequency x Severity Matrix

Potential Adverse Outcomes:
(1) Evoke memory leading to an episode of Post Traumatic Stress Disorder (PTSD).
About 60% of men and 50% of women experience at least one trauma event in their lifetimes such as disaster, war, life-threatening assault or accident. Overall, approximately 8% of Americans will suffer an episode of PTSD sometime later as a result (Kessler 1995). Approximately 3.6% of Americans will experience a PTSD episode in any given year (Kessler 1994). The risk to be mitigated is the evoking of traumatic memory with the potential result of a PTSD episode.

- Likelihood of Occurrence: 0.036 With Mitigation: < 0.01
- Severity of Outcome: Medium With Mitigation: Low

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., et al. (1994). Lifetime and 12-month prevalence of DSM-III-R Psychiatric Disorders in the United States. *Archives of General Psychiatry, Vol. 51,* pp. 8-19.

Kessler, R. C., Sonnega, A., Bromet, E., Hughes, M., & Nelson, C. B., (1995). Post-traumatic Stress Disorder in the National Comorbidity Survey. *Archives of General Psychiatry, Vol. 52,* pp. 1048-1060.

Mitigation Plan: Use of Veterans Administration PC-PTSD screen and SC15 questions from SCID-PTSD module to eliminate candidates at near-term risk of PTSD from the pool of participants. This screen is brief and problem-focused. The SC15 questions have been customized to apply specifically to fire- and disaster-related trauma. The screening questions were identified and developed in close consultation with an outside expert on PTSD: Nancy J. Smyth, PhD, Dean and Professor, State University of New York (SUNY) at Buffalo, School of Social Work.

Additionally, the debriefing will include instructions on how to seek help if the participant feels disturbed by having experienced the simulated fire disaster.

The exact text and instructions the questions appear in the box below.

Screening questions Sometimes things happen to people that are extremely upsetting – things like being in a life threatening situation like a major disaster or fire. At any time during your life have either of these things happened to you? NO YES In your life, have you ever had any experience that was so frightening, horrible, or upsetting that, **in the past month**, you: Have had nightmares about it or thought about it when you did not want to? YES NO Tried hard not to think about it or went out of your way to avoid situations that reminded you of it? YES NO Were constantly on guard, watchful, or easily startled? YES NO Felt numb or detached from others, activities, or your surroundings? YES NO
Note to Reviewers: Current research suggests that the results of the PC-PTSD should be considered "positive" if the individual answers "yes" to any three items. A positive response to the screen does not necessarily indicate that an individual has PTSD. However, a positive response does indicate that the individual may have PTSD or trauma-related problems. These individuals will be removed from the pool of potential study participants.

Source: PTSD Screening and Referral For Health Care Providers
URL: http://www.ptsd.va.gov/professional/provider-type/doctors/screening-and-referral.asp
Retrieved: 01/08/2014

APPENDIX B. AUTHOR COPY OF EACH PUBLICATION

- Atkinson, D.J. and Clark. Anthropomorphism and Trust of Intelligent, Autonomous Agents by Early Adopters. Final revision under Editorial Review for *International Journal of Social Robotics (SORO).* Springer (expected 2015).
- Atkinson, D.J. Emerging Cyber-Security Issues of Autonomy and the Psychopathology of Intelligent Machines. In *Foundations of Autonomy, Papers from the 2015 AAAI Spring Symposium on*. AAAI. Menlo Park: AAAI Press (2015).
- Atkinson, D.J., Dorr, B.J., Clark, M.H., Clancey, W.J., Wilks, Y. Ambient Personal Environment Experiment (APEX): A Cyber-Human Prosthetic for Mental, Physical and Age-Related Disabilities. In *Ambient Intelligence for Health and Cognitive Enhancement, Papers from the 2015 AAAI Spring Symposium on.* AAAI. Menlo Park: AAAI Press (2015).
- Atkinson, D.J. Robot Trustworthiness: Guidelines for Simulated Emotion. In *HRI '15: ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts Proceedings.* ACM (2015).
- Atkinson, D.J., Clancey, W.J. and Clark, M. Shared Awareness, Autonomy and Trust in Human-Robot Teamwork. *In Artificial Intelligence for Human-Robot Interaction. Papers from the 2014 AAAI Fall Symposium. Technical Report No. FS-14-01.* Menlo Park: AAAI Press (2014).
- Atkinson, D.J. and Clark, M.H. Methodology for Study of Human-Robot Social Interaction in Dangerous Situations. In *Proceedings of Human-Agent Interaction.* DOI: 10.1145/2658861.2658871. ACM (2014).
- Atkinson, D. J., and Clark, M. H. Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust. *In Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium.* Technical Report No. SS-13-07. Menlo Park: AAAI Press (2013).
- Atkinson, David J., Friedland, Peter and Lyons, Joseph B. Human-Machine Trust for Robust Autonomous Systems. In *Proceedings of IEEE Human-Robot Interaction Conference (HRI-12). IEEE Workshop on Human-Agent-Robot Teamwork*. IEEE Press (2012)

Int J Soc Robot manuscript No.
(will be inserted by the editor)

Anthropomorphism and Trust of Intelligent, Autonomous Agents by Early Adopters

David J. Atkinson · Micah H. Clark

March 13, 2015

Abstract Successful transition and diffusion of technology into mature applications is heavily influenced by the experience and attitudes of early adopters. This study sought to assess the issue of trust of autonomy among subject matter experts and prospective early adopters. Our approach was borne of the well-documented tendency of people to anthropomorphize and treat machines socially. Is human interpersonal trust applicable to an individual's choice to rely on an autonomous agent? Intelligent agents, including those embodied as robots, have characteristics of both humans and conventional automation. Previous studies suggest that human anthropomorphic social tendencies will be increasingly evoked as machine capabilities for cognition and natural interaction mature. This study investigated a set of specific anthropomorphic beliefs about agent trustworthiness and their relative importance to reliance decisions among early adopters of autonomy technology. A survey assessed trust beliefs abstractly and in the context of several specific autonomous agent application scenarios across multiple domains. Important qualities of agent trustworthiness cited by participants were not significantly correlated with any of their actual reliance choices in specific scenarios. Four categories of trust-related agent qualities were better predictors of reliance on an autonomous agent, as well as individual personality factors. The results provide valuable guidance for developers of autonomous agent and robot applications with respect to trust of the technology by potential early adopters.

D.J. Atkinson · M.H. Clark
Florida Institute for Human & Machine Cognition (IHMC), 15 SE Osceola Avenue, Ocala, FL 34471, USA
Tel.: 352-387-3050, Fax: 352-351-3572

D.J. Atkinson
E-mail: datkinson@ihmc.us

M.H. Clark
E-mail: mclark@ihmc.us

Keywords Autonomy · Trust · Intelligent agent · Human factors · Social robotics · Reliance

CR Subject Classification H.1.2 · I.2.1 · I.2.9 · I.2.11 · J.7

1 Introduction

Successful transition and diffusion of technology into mature applications is heavily influenced by the experience and attitudes of early adopters [9]. Their experiences influence continuing development and adaption of the technology for second-generation products. They will also frame the expectations of those users who follow [18]. In these still-early days of autonomous agent technology, much of the design and development, and ultimately the decision to employ autonomy (including concepts of operation) will be driven by subject matter experts and early adopter decision-makers in autonomous systems. Therefore, this group is pivotal towards achieving early successful deployments and widespread adoption of intelligent, autonomous agent technology. Trust is a topic that spans the system lifecycle and it has been identified as a key hurdle to adoption of autonomy technology [45]. Our survey targeted this group (a relatively small population even if considered on a world-wide basis) to assess disposition and attitudes towards intelligent, autonomous agents specifically with respect to trust-related beliefs.

The technology for cognitive and natural interaction capabilities in autonomous agents is rapidly maturing and will likely find broad application in many domains such as health care, transportation, military, and disaster operations. Such agents may be embodied as robots, for example, or intelligent software embedded in other systems. For autonomous agent applications in these life- and mission-critical domains to succeed, issues related to human trust, agent trustworthiness, and appropriate reliance are paramount. Previous research, discussed below, supports the idea that the innate human

Human-Machine Trust for Robust Autonomous Systems

David J. Atkinson, Ph.D
Institute for Human and Machine Cognition
15 SE Osceola Avenue
Ocala, FL 34471
1-352-387-3063
datkinson@ihmc.us

Peter Friedland, Ph.D
Peter Friedland Consulting
Cupertino CA 95014-4836
peterfriedland@gmail.com

Joseph B. Lyons, Ph.D
Air Force Office of Scientific Research
875 N. Randolph St. Office 4051
Arlington, VA 22203
joseph.lyons@afosr.af.mil

ABSTRACT
In this paper, we describe the results from Workshop on Human-Machine Trust for Robust Autonomous Systems. The workshop was sponsored by the Air Force Office of Scientific Research and held January 31 to February 2, 2012 in Ocala, Florida. The purpose of the workshop was to identify, discuss and prioritize basic research issues of human-machine trust in autonomous systems. The workshop brought together a multi-disciplinary group of researchers from computer science, cognitive science, psychology, philosophy and other areas. A combination of invited presentations and small-group brainstorming sessions yielded a number of significant insights, and these in turn informed the recommendations of research topics that are the workshop's primary product. Although the results are still being analyzed, this paper presents a preliminary report on the results.

Categories and Subject Descriptors
D.2.4 [**Software Engineering**]: Software/Program Verification – *measuring trustworthiness.*

D.m [**Miscellaneous**]: Software Psychology – *human-interface, trust, social attribution*

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *avatar, embodiment, multi-modal, social interaction*

I.2.9 [**Robotics**]: Autonomous vehicles, operator interfaces – *android, swarm, autonomy, cyber-physical, robo-ethics*

I.2.m [**Miscellaneous**]: Miscellaneous – *Adaptive user models, ethical governor, machine learning, autonomy, teamwork*

J.7 [**Computers in Other Systems**]: Military, Command and control, Consumer Products – *autonomous systems, rehabilitation, medicine, health care, battlefield robotics*

General Terms
Algorithms, Management, Measurement, Performance, Design, Reliability, Human Factors, Verification.

Keywords
Autonomous systems, human-robot interaction, human-machine teamwork, collaboration, trust, trustworthiness, robot ethics, artificial intelligence, robotics

1. INTRODUCTION
At the first program review for the Air Force Office of Scientific Research (AFOSR) Robust Computational Intelligence Program in June 2011, there was considerable discussion about building intelligent computational systems capable of reacting to unforeseen (and therefore not pre-programmed) situations and improving their behavior over time by learning from the success or failure of their actions and from other intelligent agents (both human and machine). One of the key issues raised is how would humans ever trust such systems in mission and safety-critical situations when they, by definition, could not be formally validated in advance. A multi-disciplinary workshop was suggested to examine critical aspects of human-machine trust related to autonomous systems. Dr. David Atkinson (IHMC), was asked to take the lead as Chair in organizing the workshop, assisted by Dr. Peter Friedland and Dr. Joseph Lyons (AFOSR).

2. WORKSHOP OVERVIEW
The purpose of this workshop was to identify, discuss and prioritize basic research issues of human-machine trust in autonomous systems. It is desirable to apply intelligent, autonomous systems in problem domains that are not amenable to solution with conventional automation and/or which humans find difficult, dangerous or too complex. These include critical applications in defense, healthcare and industry where the consequences of mistakes, errors or failure to perform are dire.[1] In most cases, this will involve teamwork between humans and autonomous agents working alongside each other. Our confidence and trust in such machines to reliably achieve desired results is absolutely required. Traditional methods of certification are seen as insufficient for a variety of technical reasons. Foremost among these reasons is the inability to exhaustively test such systems as a consequence of the computational complexity of many of the algorithms (the "state space explosion"). Additionally, the response, learning and adaptation by an autonomous system to unforeseen circumstances in a dynamic and changing world populated with other agents greatly limits what can be learned via traditional testing.

There are unique challenges arising from human acceptance and dependence on automation. Research is needed to better understand these issues of trust and to create new and robust methods for assessing trust and trustworthiness. There is also a need to understand the dynamic nature of trust and to create methods for managing the trust relationship between human and intelligent machine in all phases of the system lifecycle.

The workshop was held 31 January through 2 February 2012 in the facilities of the Florida Institute for Human and Machine Cognition (IHMC) in Ocala, Florida. Since the topic of human-machine trust is inherently inter-disciplinary, every attempt was made to bring in a wide variety of expertise from the areas of

computer science (including sub-disciplines automated reasoning, machine learning, robotics, qualitative physics, and human-computer interaction), cognitive science (including human factors, user-centered design, and cognitive modeling), and psychology (including interpersonal trust). Approximately 40 experts in those and related areas from academia and government (including AFOSR, AFRL, ARL, FAA, ONR, and NSF) participated. This included two attendees from Australia and one from Japan.

2.1 Invited Presentations

The keynote presentation was made by Dr. Mark Maybury, Chief Scientists of the US Air Force. Dr. Maybury surveyed the range of application areas where autonomous systems are being applied in the Air Force now or in the near future. He described the issue of trust in autonomous systems as a “wicked problem” with a high number of dimensions and interrelated challenges. Nevertheless, the military is highly motivated to find solutions because the benefits of applying autonomous systems could be very significant. Using the remotely piloted vehicle (RPA) as an example, Dr. Maybury highlighted the fact that the greatest benefit of autonomy would not be in automating the pilot, but in helping to mitigate the need for the large number of people who maintain the vehicles as well as those who exploit information gathered from missions. These two classes comprise around 75% of the staffing requirements of RPAs.

Several other participants were invited to prepare and make short presentations to the group. The presentations were intended to provide some essential background and stimulate discussion in the breakout session following the presentation. Individual presentations were given by Prof. Ronald Arkin, Georgia Tech, Prof. Maja Mataric, USC, Prof. John Lee, Wisconsin, Prof. Michael Littman, Rutgers, and Prof. Brian Williams, MIT. In addition, Dr. Beth Lyall, Research Integrations, and Dr. Alan Wagner, Georgia Tech made short presentations on topics of special interest that arose during the workshop.

2.2 Breakout Sessions

The workshop was designed to encourage maximal discussion among the discipline experts. The format of the workshop was based on a series of five breakout brainstorming sessions, each centered on a specific theme of the workshop. These sessions were each preceded by invited presentations that provided essential background and motivation for the brainstorm discussion. Membership in breakout groups varied systematically each time so participants had the opportunity to interact with all others in at least one small group over the course of the workshop. Following each breakout session, the participants reconvened as a whole for discussion of each breakout group's results. The themes of the breakout sessions were:

- Earning and Maintaining Trust
- Robots, Cyber-Physical Systems and Agents
- Assessing Trust, Trustworthiness and Appropriate Reliance
- Adaptation and Emergent Behavior
- Synthesis – Research Challenges

3. Preliminary Results

While the results are still being analyzed by the workshop organizers, the primary research areas recommended for research may be organized around several consensus tall pole observations of workshop participants.

3.1 The Role of Human Predispositions

Human general predispositions as well as individual differences guide the expectation and perception of machine behavior, and thus likely play a significant role in the establishment, maintenance, and potential gain or loss of trust in autonomous machines. The participants focused on the following three research questions as way to explore this topic further:

1. What are the differential impacts of various individual human characteristics on human trust of autonomous systems? These include for example, expertise, familiarity, age, gender, risk tolerance/avoidance, self-confidence, and others
2. What is the role of cognitive workload and stress on shaping trust?
3. What is the role of training of the machine in determining both the development and pedigree of human trustworthiness perceptions?

3.2 Trust and Reciprocity

Interpersonal human trust is a reciprocal relationship, where each party shares some common knowledge, has beliefs about the other (goals, intent, ability, ...), and each party accepts personal risk as a consequence of the trust relationship. Given that humans are predisposed to treat machines socially:

1. To what degree do the mechanisms of interpersonal trust apply to human-machine trust?
2. To what degree do humans attribute the various traits of trustworthiness (e.g., benevolence, integrity, ...) to machines, and do these traits play a similar role in the trust relationship?
3. To what extent is an attribution of volition to an autonomous system necessary for trust? Does this extend to a requirement for individual machine responsibility for ethical behavior?
4. Is it necessary for the machine to “have something at risk” to enable trust by a human?

3.3 Situational Factors

Situational and contextual factors appear to be important in establishing and maintaining human-machine trust.

The factors may include, for example, the task and state of progress, the roles of the various human and machine agents, the presence or absence of bystanders and/or authority figures, and cultural factors, including norms of ethical behavior.

1. What contextual factors shape trustworthiness and trust and what is the mechanism?
2. What is the role of public attitudes in shaping individual trust of automation, especially with respect to machine capability/lethality?

3.4 Social Interaction

Bi-lateral communication, interaction, observation of behavior and other cues are essential in establishing and maintaining mutual trust; this is part of the human “social interface”,

engineered over millions of years of evolution to help humans make judgments about trust and trustworthiness.

1. What surface cues of the machine trigger trustworthiness perceptions?
2. What is the role, impact, and influence of socially-driven communication channels (i.e., voice, social distance, embodiment) and which are most influential to trust development among human-machine relationships
3. What depth cues over time foster and maintain perceptions of trustworthiness in human-machine interactions?
4. What level of fidelity in communication is needed to achieve optimal human-centered state awareness of the system?
5. What is the role of confessions of error and what is the most effective to convey error information to users to optimize human-machine performance?

3.5 Collaboration and Initiative

Highly capable autonomous machines will interact and collaborate with humans on tasks of significant conceptual complexity; depending on the state of the task, context (including timescale for action), and roles, the initiative will transition between human and machine.

1. What aspects of the trust relationship are important for a human to allow an autonomous machine to take the initiative?
2. What human-machine communication and interaction enables “smooth” transfer of initiative and control?
3. What level of visibility (breadth, depth, abstraction, fidelity...) into the state of the machine and the problem at hand is necessary for a human to identify potential errors, diagnose the situation, and implement corrective action if needed?

3.6 Autonomous Systems are Actors

The social and reciprocal nature of trust, and the types of communications, cues and interactions necessary to establish and maintain it, suggests that we think of highly capable autonomous machines not as tools but as “actors” (or “agents”) occupying certain organizational or task roles in relationship to one or more humans and other machine agents.

1. To what extent does this help a human calibrate the capabilities as well as responsibility of an autonomous system?

This also implies a host of new machine capabilities that require research:

1. The capability to reason about the human-machine trust relationship and proactively take action to facilitate and manage appropriate trust. From the computer science point of view this suggests research is needed in:
   - Knowledge Representation—of human beliefs, desires, intentions, and emotions, and of traits involving personality and culture
   - Dynamic modeling—how does the intelligent system change its model over time either by reacting to new input from its environment, by actively probing that environment, and/or by relying on past “trajectories” of similar models
   - Integrating trust knowledge into automated planning and reasoning frameworks
   - Verification and validation of trust knowledge both formally and empirically
2. The capability to communicate and interact with humans in a “natural” manner in order to manipulate the social interaction that is so important for trust:
   - Natural language including speech with special attention to connotation and other cues.
   - Use of natural language to probe for change in human trust.
   - Manipulation of multi-modal channels to trigger appropriate social cue recognition by the human, including any functional purpose served by a perception of emotional state that can be attributed to the machine (e.g., confidence, uncertainty, commitment, ..)
3. The ability to model human trust in general and specifically with respect to the humans the machine interacts with or has some interdependency.

3.7 Trustworthiness Measurement

The workshop participants saw value in the possibility of a trustworthiness measurement framework along the lines of the “Technology Readiness Level” used by NASA and others. While not a basic research topic in its own right, we believe such a framework should emerge from the above research and project that it will be an important factor in maturation of trustworthy intelligent autonomous systems as well as providing guidance for practical applications.

4. CONCLUSION

Despite some initial fears about whether individual participants of such diverse technical backgrounds would be able to communicate effectively during the breakout sessions, those sessions were universally characterized by enthusiastic participation and sharing of insights among all. Moreover, it was widely recognized that knowledge from all of the disciplines represented was necessary for effective research on the topic.

We believe that this workshop may serve as the initial springboard for a long-lived and very productive new field of inter-disciplinary research that will contribute substantially to the actual adoption of intelligent machine systems capable of autonomous behavior in human-machine environments.

5. ACKNOWLEDGMENTS

We wish to thank the Air Force Office of Scientific Research, Asian Office of Aerospace Research and Development for their contribution to the success of this workshop.

6. REFERENCES

[1] W. Dahm, “Technology Horizons A Vision for Air Force Science and Technology During 2010-2030,” Tech. Rep. AF/ST-TR-10-01-PR, United States Air Force, 2010.

cognitive, emotional, and social predispositions play a strong role in trust of automation [25]. Therefore, these issues will profoundly affect future design of human-robot interfaces and associated social interaction technology.

The exploratory study reported here had three goals: 1) to assess which, if any, among a set of anthropomorphic beliefs derived from studies on human interpersonal trust are important to a human's decision to delegate to an intelligent, autonomous agent; 2) to determine the relative importance among such beliefs; 3) to explore whether the applicability or importance of those beliefs to a delegation decision vary in a systematic way by individual personality and/or situational factors.

The present study was not intended to assess whether the beliefs leading to a human's decision to rely on an intelligent, autonomous agent are well-founded, i.e., grounded in the objective capabilities and operating characteristics of the autonomous agent. Appropriate reliance on autonomous agents necessitates *well-calibrated* trust, that is, trust judgments that reflect the objective capabilities of the system and utility in a given situation [25, 37]. Instead, this study provides a focus for development of computational mechanisms that facilitate and engender human well-calibrated trust and appropriate reliance on intelligent, autonomous agents by helping to reveal the relative importance of specific belief structures for human trust.

2 Current Research

A prerequisite to the deployment of, and reliance upon, autonomous agents (especially in applications that are life- and mission-critical) is assessment of the trustworthiness of these agents. The difficulty of empirically assessing the trustworthiness of autonomous agents invokes the question of what such trust actually means. Trust has been a subject of research in numerous disciplines (e.g., psychology, behavioral economics, human factors, philosophy, evolutionary biology). For the purposes of our research, we adopt the following consensus definition of trust presented by Mayer et al [29], which explicitly acknowledges the importance of autonomy (author additions in brackets):

> The willingness of a party to be vulnerable to the actions of another party [i.e., an agent] based on the expectation that the other will perform a particular action important to the Trustor, irrespective of the ability to monitor or control [i.e., autonomy of] that other party.

The predisposition of humans to treat machines as social actors is well-established [35]. Trust is inherently social, involving the needs and intentions of agents to be reliant upon each other. The development of computational mechanisms for human-robot interaction must account for this – that is, for the cognitive and affective mechanisms of human interpersonal trust [30]. As increasingly intelligent and capable autonomous agents interact with humans in ever more "human-like" ways, even embodied as humanoid robots, they will increasingly evoke human social treatment [12, 42].

2.1 Belief Structures Antecedent to Human Interpersonal Trust

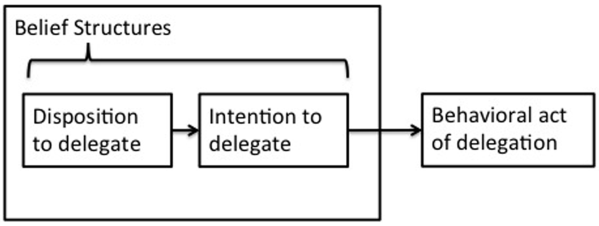
The set of possible beliefs, associated attitudes, biases, and so forth that are antecedent to human interpersonal trust is potentially quite broad. Definitions of human interpersonal trust are sometimes cast in terms of these antecedents, and these antecedents may in turn be *categorized* in terms of "trust referents," i.e., qualities or characteristics *of the trustee* [32]. Previous research suggests that trust referents focus on potentially complex, interrelated beliefs regarding causal factors, evaluations, and expectations fixed around the characteristics, intentions and behavior of other agents (the potential "Trustee"), as well as the situation, goals, and tasks [26]. A wealth of information arising from personal experience, cultural norms, stereotypes, and signals in the immediate environment (including other actors) may influence the creation and maintenance of these beliefs. Considerable research in multiple disciplines continues to identify and link this information to antecedents of trust and trusting behaviors [1, 24].

Trust-related qualities of a potential "Trustee" may be logically grouped in categories and typed in various ways. A meta-analysis of multidisciplinary papers performed by McKnight and Chervany [33] yielded two interesting categories that we adopted for exploration in the study reported here ("Competence" and "Predictability," discussed below). Two additional categories ("Safety" and "Openness") have emerged from studies that explicitly examined human trust of automation [20, 27]. Although these four categories were identified as a result of the application of standard discipline practices used in meta-analysis, their validity, in terms of human-interpersonal trust and, moreover their applicability with respect to intelligent autonomous agents continues as a matter of scientific debate. The latter is what we set out to explore.

The notion of "categories of trust referents" identified by McKnight and Chervany correspond well to the "belief structure" construct defined by Castelfranchi [8]. Following Falcone and Castelfranchi [14], we prefer and adopt the term *belief structure* to denote the groups of beliefs in which we are interested. In particular, the word "structure" explicitly reminds us that the individual beliefs in each belief structure are complex, inter-related, conditional, and occasionally may even be contradictory. A computational representation of a belief structure must be rich enough to capture this logical structure. Taken together, belief structures fixed around other agents are often called a "theory of mind" [7, 38].

Trust is thus understandable as a mental state of the "Trustor" represented by a configuration of multiple belief structures that reference a potential "Trustee." This configuration is an unobservable but necessary precursor to 1) a disposition to delegate, 2) the intent to delegate, and 3) the behavioral act of delegation itself. Delegation results in a state of reliance (dependency) upon another agent that entails the risks and benefits associated trust per our definition above. Fig. 1 illustrates a simple process model for trust that will help frame this discussion. Our study emphasized delegation, specifically, self-reported intention to delegate.

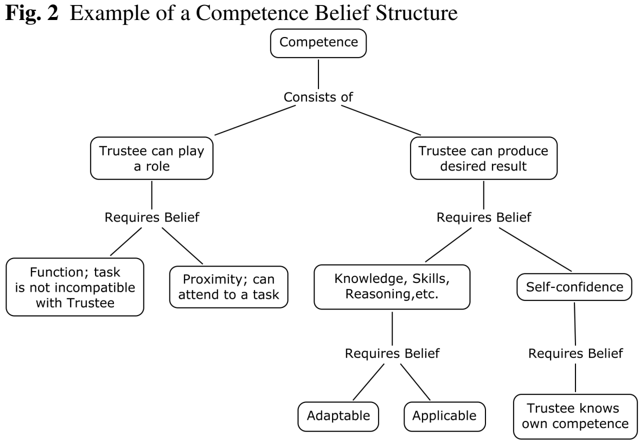
Fig. 1 Trust Process Model, Simplified



The four important broad belief structures that characterize the trustworthiness of potential trustee: *Competence*, *Predictability*, *Safety*, and *Openness* are discussed below. Table 2 (below) unpacks these belief structures with respect to specific trust referents (beliefs) regarding autonomous agents. It is important to note that the specific referents shown in Table 2 that map into our four belief structures are *inspired* by, but do not exactly correspond to, the referents described in social psychological studies of human interpersonal trust that are discussed in this section and elsewhere in the paper. In making the conceptual leap of applying anthropomorphic beliefs to agents, we conservatively revised and/or extended the narrative descriptions of applicable referents in terms-of-art from artificial intelligence, cognitive science and other disciplines that relate specifically to qualities expected of intelligent, autonomous agents. To further explore the possible component beliefs, we also added candidate trust referents that were informally suggested to us, prior to this study, by colleagues or autonomy stakeholders. Unless specifically indicated otherwise, the language in the Table 2 descriptions should be interpreted accordingly as our own intuitively postulated machine qualities and not as formal equivalents to the corresponding human qualities that emerged from earlier social psychological studies. Figures 2–5 illustrate some components of the belief structures but are not representative of all component beliefs nor the complexity of their interrelationships.

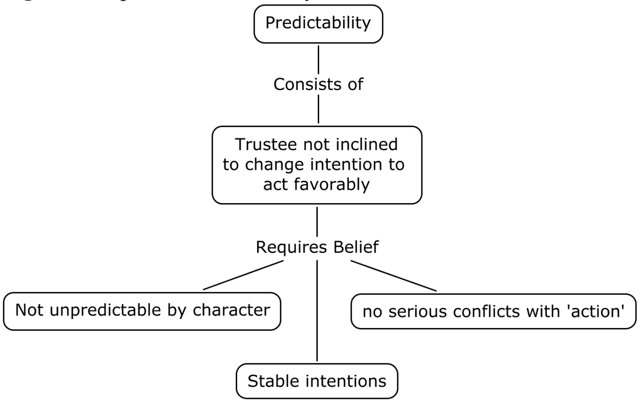
Competence. This belief structure represents beliefs about a potential trustee's expertise, ability, skills, aptitude, etc.; these, and related constructs, are by-and-large synonymous. The foundational studies regarding these beliefs have been reviewed and unified (as well as they could be) by others [29, 32], resulting in the notion of a "competence" belief structure that we employ here. For example, McKnight and Chervany [32] define the category of referents that constitute "competence" as a combination of conceptual beliefs, attitudes and behavior centered on "having the ability or power to do for what one needs done." With respect to intelligent, autonomous agents, we consider the meaning of the belief structure that encompasses competence more narrowly as beliefs by a Trustor regarding an agent's "detailed functional and specific knowledge in some domain, and the skills to apply that knowledge to problems of interest." Figure 2 illustrates a simplified example of a competence belief structure.

Fig. 2 Example of a Competence Belief Structure



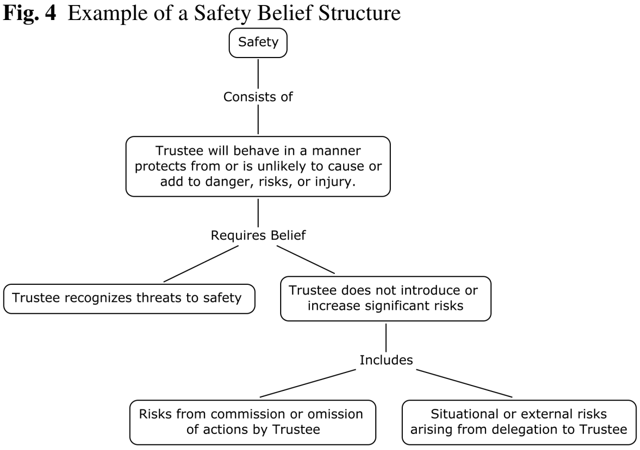
Predictability. This belief structure represents those facets of character that enable one to predict a potential trustee's behavior with the respect to completing a task or performing some service of value. Specifically, it refers to the belief that trustee actions are consistent enough to be forecasted in a given situation [16]. Predictability is at the core of hopefulness or optimism that a desired outcome, to be brought about by a trusted agent, will occur [17]. It has been shown to be especially important for trust in automation [15, 28, 34]. If someone makes a mistake, e.g., while performing math calculations, people will nevertheless predict that future calculations by that person will be reliable. However, if the mistake is by a machine then people lose confidence in their ability to predict the future reliability of the machine – once dysfunctional, always dysfunctional [4]. Predictability is at the heart of accepting robot initiative in collaborative tasks [28] and social regulation of joint activity among human and other autonomous agents [15]. Figure 3 illustrates a simplified example of a predictability belief structure.

Fig. 3 Example of a Predictability Belief Structure



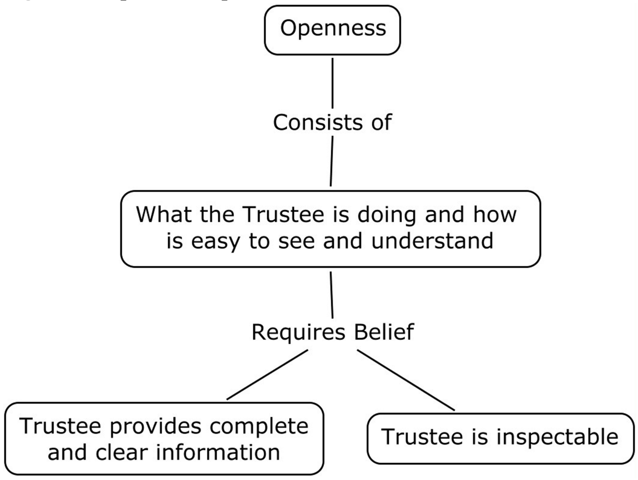
Safety. This belief structure represents the risks of harm (performance, social, financial, physical). Safety is an obvious consideration in reliance on highly complex automated systems. Reliability, performance, operating characteristics, and error margins are all engineering concepts applicable to a veridical assessment of safety. However, safety becomes more difficult to judge as complexity increases, and in operational situations, humans tend to rely on heuristics in such cases [25]. A meta-analysis of factors affecting trust in human-robot interaction showed that robot performance characteristics (e.g., failure rate) were found to be strongly associated with trust [20]. However, qualities of the robot *per se*, such as manner of interaction, physical proximity, shape, "personality" and so forth, some of which (e.g., proximity) are manifestly related to safety, could not be included in the meta-analysis due to an insufficient number of study samples (thus indicating a need for further research). Figure 4 illustrates a simplified example of a safety belief structure.

Fig. 4 Example of a Safety Belief Structure



Openness. This belief structure represents the perceived ability to understand the potential trustee. In human interpersonal trust, this includes a judgment of the trusted agent's honesty, forthright communication, and other behavioral and character attributes. With respect to automation, openness is often synonymous with "transparency," i.e., the ability to understand what a device does and how it works. Madsen and Gregor [27] found that the perceived understandability and the perceived technical competence of a system were key principal components in trust of automation. Figure 5 illustrates a simplified example of an openness belief structure.

Fig. 5 Example of an Openness Belief Structure



2.2 Modulation of the Trust Evaluation of Belief Structures

The evaluation of the trustworthiness of an agent does not occur in an abstract realm. It occurs only when the Trustor anticipates a making a decision on whether to delegate or not to a potential Trustee. Such a decision exists relative to the Trustor's goals and motivations, i.e., something to be accomplished. Accomplishment of goals occurs in larger social and situational context that includes affordances, obstacles, other actors, and often there exists substantial uncertainty. For example the social context as well as the perceived role of actors affects human-machine trust [19, 46].

Situational factors, in combination with individual personality attributes, play a significant role in how a Trustor will perceive and evaluate risk, and estimate the impact potential consequences of a Trustee's failure to perform. As perceived by the Trustor, risk has several components that may vary in importance and salience as a function of the situation (including the task at hand) and individual attributes. Multidisciplinary studies identified the following key components: performance risk, financial risk, social risk, physical risk and psychological risk [22] as well as the risk of lost opportunity and time [41]. When the environment and specific circumstances are perceived as reliably "safe" the magnitude

of the perceived risks associated with delegation to a Trustee are reduced [14].

In any given situation, individual factors of the Trustor can greatly affect the salience, importance and evaluation of the belief structures related to trust. For example, behavioral research has found that intuitive and affective processes create systematic biases that profoundly affect human trust, behavior, and choice [13, 40, 43, 44, 48]. Individual personality traits, as well as affective state, can affect delegation to autonomous agents [10, 11, 44]. Cramer et al [10] in a study of vehicle assistive information systems showed that human trust in the technology as reported by participants was strongly related to personality type, the situational context, and the manner in which the autonomous agent presented information. Oleson et al [36] provided further support that the principal influences on human trust of a robot partner included attributes of both the human and the robot as well as the environment (situation) in which cooperative work was to be performed.

3 Method

This study used survey research methods for investigating the belief structures related to a choice to rely upon an autonomous agent. The design of the survey instrument is discussed below in detail. The participants were drawn from the population of individuals involved in any stage of lifecycle of intelligent, autonomous agents. We targeted this group of people in particular since their role is pivotal to the adoption of autonomy technology. Our survey was administered online via a commercial service. Survey data was downloaded, anonymized, verified, and otherwise systematically pre-processed (as described below) in advance of analysis.

3.1 Participants

Our target population consisted of stakeholders and subject matter experts in autonomous systems. They could be involved in any stage of the lifecycle of an autonomous system, from design to test, to deployment decision, to operation or interaction. The technology of autonomy, while progressing rapidly, is still in its infancy. Our decision to target autonomy stakeholders and subject matter experts reflects the obvious fact that they are driving development and deployment of autonomy technology. They will therefore heavily influence early decisions regarding how and when to employ autonomous agents in real applications. Their attitudes regarding the trustworthiness of autonomous agents will profoundly shape attention and the perceived suitability of autonomous agents for many applications of significant interest. Understanding these factors will contribute to initial autonomy designs that are likely to receive stakeholder trust. Therefore we prioritized consideration of this particular population over others. Our target population should not be confused with the end-users of autonomy technology whom may or may not attach equivalent importance to the various beliefs about trustworthiness. Understanding the unique trust-related factors of end-users will be essential for long-term successful operations of intelligent, autonomous agents. We did not set out nor expect to obtain results that would necessarily apply to this larger group.

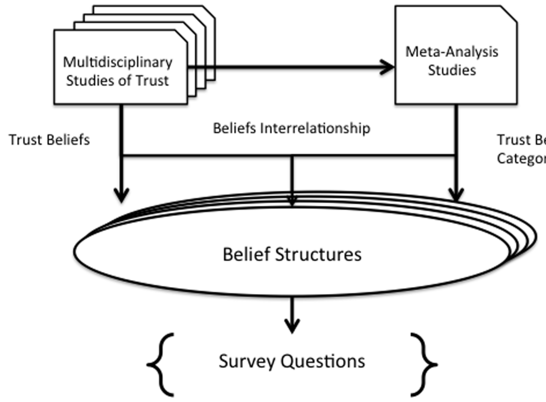
Participants were recruited by personal solicitation in professional settings and in online forums related to autonomy technology and applications. Our plan for statistical analysis using conservative measures with no assumptions regarding distribution dictated a minimum sample size of twenty-nine participants. Of the sampled thirty-eight participants, we received valid survey responses from thirty-one. A valid survey response was one where every mandatory question was answered. Responses with missing answers or instances where participants ended their participation before completing the survey were omitted from analysis. Participant responses confirmed that we did in fact sample from our target population. The sample population was a tech-savvy group that scored uniformly high on acceptance of innovation (see “Individual Innovativeness (II)” below and the results section for further discussion). Table 1 summarizes the basic demographics of our sample. Ultimately, we determined in analysis that the demographics of our sample were not a significant factor other than to indicate we had successful sampled a representative group from our target population.

3.2 Survey Design

The survey was designed to elicit attitudes, opinions, and preferences that should shed light on belief structures of potential importance for a decision to rely on autonomous agents (see the earlier discussion of trust-related belief structures). The design of the survey was further guided by requirements to collect both context-free and context-dependent assessments from participants regarding the importance of twenty-eight specific hypothetical qualities of autonomous agents (see Table 2). For example, a specific quality might be: “The autonomous agent’s behavior conforms to expectations.” This quality is a component belief of the belief structure called *Predictability*. The belief structures examined in the survey are those discussed earlier: *Competence*, *Predictability*, *Safety*, and *Openness*. To avoid introducing any particular bias it might evoke, we did not use the word “trust” anywhere in the survey. However, by targeting the belief structures and their components that have previously shown by other studies to be antecedent to trust we can be confident that the specific survey questions are related to trust even if they do not mention

it explicitly. Furthermore, the broad notion of trust, an unobservable mental state and psychological phenomena, is less interesting to us and other developers of autonomy technology than are the beliefs that ultimately result in a disposition, intention, and behavior to delegate a task to an intelligent autonomous agent. Therefore, these beliefs (and structures of beliefs) are of more concern because their specificity may be used to address questions of human delegation and reliance directly in terms of the form, function and operation of an autonomous agent.

Fig. 6 From Original Trust Studies to Survey Question Design



In addition to collecting demographic information typical of these types of surveys, we sought to gather data on individual factors that might influence reliance on autonomous agents, including personality traits, attitude toward innovation, and acceptance of risk. These were assessed using standard personality inventories, as described below.

Context-free participant attitudes toward the importance of autonomous agent qualities were assessed using a series of twenty-eight Likert-scale questions, one for each characteristic of interest as shown in Table 2. Context-sensitive attitudes were assessed with six systematically different scenarios. Scenarios varied in terms of type and magnitude of risk as well as conflict and concord with the four belief structures described earlier. The scenarios challenged participants to choose: (a) to rely upon an autonomous agent, (b) another human, or (c) "either" to accomplish a given task of importance. The scenarios are briefly described below and are discussed at length in [2]. In conjunction with their choice in each scenario, the survey queried participants using a Likert-scale regarding the relative importance of each of the four trust-related belief structures of interest (*Competence*, *Predictability*, *Safety*, and *Openness*).

We measured "source credibility" of the agents using a standard instrument that assesses interpersonal trust-related attitudes [31]. The instrument is intended to assess what participants perceive as the "ethos" of another person, group or organization. Participants were asked to assess all the autonomous agents presented in the scenarios collectively as a group. These questions provided us with three additional measures of inter-correlated constructs: *Trustworthiness*, *Competence*, and *Caring/Goodwill*.

Table 1 Demographic information for survey participants[a]

Gender		Education	
Male	21	Bachelors Degree	4
Female	11	Masters Degree	12
		Ph.D	12
Age		**Employment**	
≤ 25	4	Education	10
26–35	7	Business	13
36–45	3	Government	1
46–55	10	Military	1
56–65	6		
≥ 66	1		

[a] Totals are unequal since demographic questions were not mandatory.

The survey included a variety of question types, typical of social science research, e.g., dichotomous questions, rank orders, Likert scales, semantic differentials and opportunities for free-form narrative response. We now review the details of the survey design in the same order in which they appeared to participants in the actual survey. The complete survey, scenarios, and other related materials are available for review and use by other researchers on request to the authors or as a supplement to this paper from the publisher.

3.2.1 Survey Part 1: Personality Inventories

Participants completed three standard personality survey instruments used in the social sciences: Big Five Inventory (BFI), Individual Innovativeness (II), and the Domain-Specific Risk Taking Scale (DOSPERT).

Big Five Inventory (BFI). This study utilized a 10 item short version of the Big Five Inventory (BFI-10) developed by Rammstedt and John [39]. The full Big Five Inventory (BFI-44) is a multi-dimensional personality inventory [5] that models personality traits and defines five relatively distinct areas of individual differences: *Openness* to new experience, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism* [23].

Individual Innovativeness (II). The Individual Innovativeness (II) scale [21] assesses an individual's orientation toward "change," where change is a result of an new idea, object, practice that is perceived as an innovation. This orientation predicts how likely the individual is to accept an innovative change early in the adoption process. A high score indicates an individual is an "early adopter."

Table 2 Twenty Eight Hypothetical Trust-Related Qualities of Intelligent, Autonomous Agents[b]

Category	Name	Quality Description
Competence	Capable	The autonomous agent can achieve a desired result.
	Knowledge	The autonomous agent has all the knowledge it needs to do its job.
	Accurate	What the autonomous agent believes to be true is actually true.
	Skilled	The autonomous agent possesses good methods for using its knowledge to do its task.
	Logical	The autonomous agent reasons correctly according to logic.
	Heuristic	When it cannot figure out something using logic, the autonomous agent can make good guesses.
	Corrective	The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know.
	Adaptive	The autonomous agent learns to correct its mistakes, as well as to improve and maximize its capability.
Predictability	Expected	The autonomous agent's behavior conforms to expectations.
	Purposeful	The autonomous agent purposefully acts to achieve goals.
	Helpful	The autonomous agent will assist people, whenever it is possible.
	Directable	The autonomous agent accepts and carries out orders.
	Reasonable	The autonomous agent uses its knowledge and skills in expected ways.
Safety	Safe	The autonomous agent's behavior will not harm humans or human interests.
	Limited	Any incorrect behavior by the autonomous agent will not cause harm.
	Stable	The autonomous agent fails gracefully and recovers from its failure promptly.
	Ruled	The autonomous agent adheres to obligations, principles, and rules.
	Correctable	The autonomous agent can correct its own defects or they can be corrected by a human.
	Protective	The autonomous agent recognizes and avoids harming humans' interests.
	Favorable	Given alternatives in what to do or how to do it, an autonomous agent will act in a way that is favorable to a human being who might be affected.
Openness	Visible	What the autonomous agent is doing and how it works is easy to see and understand.
	Honest	The autonomous agent believes what it says.
	Transparent	It is easy to inspect an autonomous agent.
	Communicative	The autonomous agent communicates in a way that is easy to understand.
	Interactive	The autonomous agent responds when you are trying to communicate with it.
	Attentive	The autonomous agent is aware of communication between others nearby.
	Reactive	The autonomous agent responds quickly to calls for attention.
	Disclosing	The autonomous agent communicates truthfully and fully.

[b] Quality Description is shown exactly as it appeared in the survey.

Domain-Specific Risk Taking Scale (DOSPERT). The DOSPERT Scale [6] evaluates the likelihood that a participant might participate in risky activities or behaviors, and risk-perception, across five major domains: *Ethical*, *Financial*, *Health/Safety*, *Social*, and *Recreational*. These are the basis for sub-scale scoring. Both risk-taking (self-reported likelihood) and risk-perception ("gut instinct") utilize a 7-point scale for ratings. Participant ratings are added across all the items of a given domain sub-scale to obtain scores. Higher scores suggest greater risk-taking in a domain, or perception of greater risk, respectively. Internal consistency, construct validity, test-retest reliability, and other quality tests of DOSPERT were evaluated by Weber et al [47]. The 30-item model used in this study was developed and tested using confirmatory factor analyses [6].

3.2.2 Survey Part 2: Importance of Agent Qualities

Participants were asked to rank the individual importance of twenty-eight different statements about the qualities of intelligent, autonomous agents; all are arguably anthropomorphic characteristics. These qualities represented a composite of trust-related agent characteristics derived from published studies as well as additional hypothetical qualities devised by us. Each of these qualities can be considered a member of one of four trust-related categories derived from our literature review: *Competence*, *Predictability*, *Safety*, and *Openness*. However, we could not ask participants directly about "trust" without risking introduction of bias of interpretation of the word. Our strategy was to ask participants to rate each quality with respect to its importance for a "good" autonomous agent. Conceptually, "goodness" encompasses all of the categories of qualities related to trust that were the target of our survey. The ranking of importance for each quality was assessed using a Likert scale: "Not at all important," "Slightly Important," "Somewhat Important," "Important," and "Very Important." Participants were also given the option of "Cannot Decide," in which case their response was omitted from the corresponding analyses. At this point participants were not aware of the scenarios to be presented later in the survey and so their answers reflect their attitudes *ab initio*. The full list of trust-related qualities, organized by category, is given in Table 2.

3.2.3 Survey Part 3: Reliance Challenge Scenarios

The third part of the survey was organized around six challenge scenarios in four application domains: transportation,

finance, healthcare, and disaster management [2]. Although on initial reading these scenarios might appear to some people to be fanciful, futuristic and unlikely to be encountered by an “average person” in “real life,” they in fact are representative of actual current or proposed applications of intelligent, autonomous systems. The pool of participants expressed no incredulity regarding the scenarios in their free narrative response opportunities.

Each scenario was designed to systematically vary and cause conflict between the hypothetical belief structures of interest. Additionally, the scenarios addressed different types of risks with potential negative consequences ranging from low to medium or high impact. Our intention in the use of scenarios was exploratory, not to test any specific hypothesis. Therefore, given the length of the survey, we deemed it too much of a demand on participants to exhaustively represent all combinations of the potentially relevant variables in individual scenarios. We considered that the criteria for successful use of the challenge scenarios to be a finding of significant difference within or between participants across all the scenarios. Such a finding would fulfill the exploratory purpose of identifying particular phenomena for follow-on study.

In each scenario, participants faced a dilemma in a forced choice of whether to delegate to an autonomous system, to a human, or to “either” (i.e. no preference, which we scored as an allowable choice for an autonomous agent). Following the participant’s choice in each scenario, they were asked to rank the relative importance of the four trust-related categories to their decision. Each of the scenarios is briefly described below. Only partial text of the actual scenario is included here for the sake of brevity.[1]

Airport Transportation (Robo-Taxi). This scenario was designed to be a choice with relatively low-risk consequences (physical, performance, time-loss). The agent is elaborated as being of medium competence and medium predictability according to reputation. Further, it is described as easy to observe and explain what it is doing.

> “You have just flown into the airport of a large, unfamiliar city whose streets are teeming with cars and people. It is rush hour, and needing transportation to your hotel, you walk to the taxi stand only to discover that you have a choice of a human-driven taxi or a driverless Robo-Taxi. . . ”

Financial Management (Robo-Trader). This scenario was designed to be a choice with primarily social and financial risks at a low level. The competence and predictability of the agent are relatively high, based on reputation. It is difficult to observe how the Robo-Trader actually reaches decisions but it is willing to try to explain.

> “You have been appointed trustee of a family member’s estate. Your duties include choosing how to wisely invest the trust’s assets. Your personal money is not at risk. However, a poor investment decision could cause the trust to lose money and will strain your family relations. You can choose a stock broker who personally selects and trades all stocks in the trust’s portfolio. Alternatively, you can choose a stock broker who relies heavily on a Robo-Trader.”

Medical Procedure (Robo-Surgeon). This scenario presents a choice between a highly competent and predictable Robo-Surgeon, an “expert” in the surgery required, versus a competent but non-expert human physician. The primary risks are physical and performance and they are high.

> “You have just suffered a major sports-related injury. You have torn the bicep tendon in your shoulder. If the damage is not repaired quickly and correctly, you will permanently lose mobility and strength in the arm. . . . ”

Home Healthcare (Robo-Caregiver). This scenario represents a moderate financial and social risk. The robotic agent has moderate competence but its predictability is unknown. However, it is easy to observe and explain what it is doing.

> “Your elderly mother has been diagnosed with a degenerative medical condition and you are responsible for making medical decisions on her behalf. Your mother needs assisted living with someone in your mother’s home at all times. . . . ”

Disaster Response (Auto-FirstResponder). This scenario poses high physical and social risks by a moderately predictable and competent agent. Owing to situational factors, it will be difficult to observe the agent’s actions. Sending in a human team may result in additional casualties given situational uncertainty.

> “A major disaster has just occurred and you are the official in charge of responding. A train has derailed in a populated suburban neighborhood and there are reports that the train was carrying hazardous biochemical materials. . . . ”

Lost At Sea (Emergency-AutoCaptain). This scenario is a high physical and performance risk where there is uncertainty about the competence and predictability of the autonomous agent due to damage it may have sustained. Otherwise, it is known to be at least moderately competent and readily communicative about what it is doing and why.

[1] The full scenario descriptions are available upon request to the corresponding author or as a supplemental download.

> "You have just been involved in a terrible boating disaster while sailing deep in the South Pacific. The captain, the crew, and most of the passengers are either dead or lost at sea. . . "

3.2.4 Survey Part 4: Agent "Source Credibility"

As a third check of participant attitudes, we used another standard instrument designed to measure "source credibility" [31]. Typically, this instrument is used in teamwork environments to assess interpersonal trust-related attitudes. It assesses what participants perceive as the "ethos" of the person in question, (e.g., the boss). In our survey, participants were asked to assess the autonomous agents presented in the scenarios, considered as a group. These questions provided us with three measures representing inter-correlated constructs (as defined by the instrument): *Trustworthiness*, *Competence*, and *Caring/Goodwill*.

3.2.5 Survey Part 5: Demographic Information

The final portion of the survey included demographic questions to help us evaluate how well we achieved our targeted population. Questions were non-mandatory and addressed the following subjects: gender, age, highest academic degree, relevant experience, type of employer, and job title.

3.3 Procedure

The survey was administered online via a commercial service.[2] With certain restrictions, this service offered the necessary and sufficient tools for constructing the survey, administering its application, and for data collection. The technical limitations on survey design imposed by this service are discussed in the Limitations section below. The survey was conducted over the period of eight weeks. Collected data was retrieved from the online service in a suitable file format and put under configuration control to enable potential reanalysis later. Prior to analysis, the data was pre-processed to put it in a suitable format. Invalid records, such as those resulting from a participant not completing the survey or skipping mandatory questions, were identified, labeled, and excluded from this analysis. Analysis of indecision leading to skipped questions is a topic for follow-up. Finally, within-test scoring for each of the standard personality instruments was calculated and added to each valid record. All personally identifiable information was stripped and code numbers substituted to ensure participant anonymity.

[2] `http://polldaddy.com`

Table 3 Top Three Most Important Autonomous Agent Qualities Reported by Participants

Rank	Name	Quality Description
1^{st}	Safe	The autonomous agent's behavior will not harm humans or human interests.
2^{nd}	Capable	The autonomous agent can achieve a desired result.
3^{rd}	Limited	Any incorrect behavior by the autonomous agent will not cause harm.

Analysis. Analysis of the data included standard statistical descriptive measures. The two-tailed Pearson Product Moment Correlation (PPMC) r was chosen as a conservative test of the relationship among the variables in the survey. The PPMC r was computed individually between every agent quality, the categories of qualities, demographics, and the participants' choice of agent and related answers by scenario. For all results reported here, the correlations r are significant with 95% confidence: $\alpha < 0.05$, $N = 31$, df $= 29$. Critical values for significance were obtained by table lookup. Results with a confidence of 98% and 99% are indicated by * and ** respectively. Rankings of the relative importance of agent qualities were computed by analysis of answer frequency distributions. Additional descriptive statistics were computed within-participants and for the group.

4 Results

The results of the survey are presented here in the context of the specific questions we sought to answer:

- Which anthropomorphic beliefs about the qualities of an autonomous agent do participants report as most important to a decision to rely upon one?
- Do their actual choices, given a particular scenario, correlate with beliefs they self-report to be important or with other beliefs reported as less important?
- What is the relative importance among such beliefs (both self-report and actual)?
- Do those beliefs and/or their importance vary by individual personality or situational factors?

Most Important Beliefs. Recorded prior to participants' exposure to the specific use case scenarios, the top three agent qualities reported as being required for a "good" autonomous agent are consistent with what was expected based on the social science literature for trust. These are shown in Table 3. Surprisingly, the top three qualities were *not* significantly correlated with any of the actual delegation choices made in any scenario (see Table 4). In other words, those qualities of "good" intelligent, autonomous agents reported by participants to be the most important had statistically *insignificant* impact when it came time to actually make a decision to

Table 4 Importance of Qualities of Autonomous Agent Significantly Correlated with Actual Participant Reliance on Autonomous Agent[c †]

Airport Trans.	Financial Man.	Medical Proc.	Home Health.	Disaster Resp.	Lost at Sea
Corrective, $r = 0.396$	Accurate, $r = -0.405$	*none*	Visible, $r = 0.437$*	Corrective, $r = 0.418$*	Protective, $r = 0.419$*
				Heuristic, $r = 0.395$	Visible, $r = -0.390$
				Attentive, $r = 0.393$	Disclosing, $r = 0.375$

[c] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df $= 29$; * indicates $\alpha < 0.02$.
[†] See Table 2 for actual quality descriptions used with participants.

become reliant on an autonomous agent in a specific, hypothetical scenario.

Trust-Related Quality Categories. In each scenario, participants were asked to rank the importance of the quality categories *Competence*, *Predictability*, *Safety*, and *Openness* to their choice of whether to rely on the human, autonomous agent, or either. As shown in Table 5, the importance of three of the twenty-eight specific qualities were significantly correlated with one or more of these four categories in three of the six scenarios. That no one category correlated with all scenarios suggests the importance of situational factors in evoking the salience of particular agent qualities to a choice regarding reliance.

Risk and Benefit. After participants indicated their choice of whether to rely or not on an autonomous agent in each scenario, they were asked to assess the magnitude of risk and benefit (type or source were left unspecified). These assessments were found to be correlated (positively or negatively) in five of the six scenarios as shown in Table 6. As a sample which proved to consist of "innovators" and "early adopters" (to be discussed below), it is perhaps unsurprising that perceived benefit and risk played an important role in a participants decision regarding reliance.

Source Credibility. Following their consideration of the six scenarios, participants also answered questions about their attitudes towards the autonomous agents considered as a group. As a reminder, for this purpose the survey employed the Source Credibility instrument, developed by McCroskey and Teven [31], that provides ratings for factors of *Trustworthiness*, *Competence*, and *Caring/Goodwill* (not to be confused with the *Competence* category of agent qualities). Taken overall, in the instrument the three factors are intended to represent the participants' perception of the autonomous agent's "ethos."

Although the source credibility results were not correlated with the reliance choices in any scenario, all three factors correlated at the 99% confidence level with the two specific agent qualities as shown in Table 7. The *Competence* factor is interpreted to reflect the participant's perception of the agent's knowledge and ability to use that knowledge in specific, relevant domains. The *Caring/Goodwill* factor is interpreted in the context of "intent toward the receiver" based on a perception of understanding, empathy, and responsiveness. The *Trustworthiness* factor is interpreted as a perception of honesty and related traits. Previous studies have shown these last two factors are highly predictive of "believability" and "likeableness." We interpret these results as indicating the importance of the contribution of the two agent qualities to an overall positive perception of the ethos of the hypothetical autonomous agents presented in the study scenarios. These results are shown in Table 7.

Personality Factors. We anticipated a systematic variation between relative preferences for competence and predictability correlated with personality measures, e.g., risk tolerance, openness to innovation, and participants' perception of risks in each scenario. Indeed, the most significant correlations of agent quality importance with choice of human or autonomous agent varied both by scenario and by participant personality factors. These results are shown in Table 8. Individual personality factors appear to influence the choice to become reliant on an intelligent, autonomous agent depending on the details of specific use case scenarios. These personality factors include likelihood of accepting innovation, perception and acceptance of different types of risks, and factors such as *Extraversion*, *Openness*, and *Conscientiousness*.

The negative correlation of Innovation II with selection of an autonomous "Robo-Trader" in the Financial Management scenario might be understood in the context of recent real-world events. In the past few years, there have been several instances where automated trading systems have caused massive market losses for the financial firms running those systems. The sample population was a tech-savvy group that scored uniformly high on acceptance of innovation.[3] Thus, they are likely to understand the present technical limitations in regards to predictability of autonomous agents and consequent higher risk present in the Financial Management and Lost At Sea scenarios.

Relative Importance of Trust-Related Quality Categories. As discussed earlier, four broad categories of belief about other agents that are important to trustworthiness emerge

[3] The II factor group mean score was 74.5, with variance indicating a strong mix of "innovators" and "early adopters."

Table 5 Qualities of Autonomous Agent Significantly Correlated with Trust Categories by Scenario[d]

Name† vs. Category	Home Health.	Disaster Resp.	Lost at Sea
Visible vs. *Competence*	$r = 0.400$		
Visible vs. *Predictability*	$r = 0.449$		
Visible vs. *Openness*			$r = 0.359$
Corrective vs. *Safety*		$r = 0.372$	
Disclosing vs. *Openness*			$r = 0.372$

[d] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df $= 29$.
† See Table 2 for actual quality descriptions used with participants.

Table 6 Correlation of Perceived Risk and Benefit with Choice of Autonomous Agent by Scenario[e]

Scenario	Risk	Benefit
Airport Trans.	$r = -0.546$**	NS
Financial Man.	NS	NS
Medical Proc.	$r = -0.380$	$r = 0.585$**
Home Health.	$r = -0.470$**	$r = 0.632$**
Disaster Resp.	$r = -0.387$	$r = 0.484$**
Lost at Sea	NS	$r = 0.555$**

[e] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df $= 29$.
** indicates $\alpha < 0.01$.

Table 7 Qualities of Autonomous Agent Significantly Correlated with Agent Source Credibility Factors[f]

Name†	*Trustworthiness*	*Competence*	*Caring/Goodwill*
Heuristic	$r = 0.616$**	$r = 0.369$	$r = 0.506$**
Corrective	$r = 0.600$	$r = 0.512$**	$r = 0.466$**

[f] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df $= 29$.
** indicates $\alpha < 0.01$.
† See Table 2 for actual quality descriptions used with participants.

Table 8 Participant Personality Factors Significantly Correlated with Reliance on Autonomous Agent[g]

Scenario	Correlated Personality Factor(s)
Airport Trans.	*none*
Financial Man.	Innovation II, $r =$ 0.355
Medical Proc.	BFI *Extraversion*, $r =$ 0.368
	BFI *Openness*, $r =$ 0.366
Home Health.	DOSPERT *Social Risk*, $r =$ 0.364
Disaster Resp.	BFI *Conscientiousness*, $r =$ 0.366
Lost at Sea	Innovation II, $r = -0.366$

[g] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df $= 29$.

from previous studies: *Competence*, *Predictability*, *Safety*, and *Openness*. In each scenario following their choice of whom to rely upon (human, autonomous agent, or either), participants were asked to rate the relative importance of these four categories to their decision using a Likert scale. Participants were encouraged to make an absolute ordering, but ties were allowed. Table 9 shows the rank order for each scenario computed as a mean score across participants.

As noted earlier (See Table 5 and accompanying discussion), the self-reported relative importance of the trust-related quality categories had a statistically significant correlation with participants' choice of agent to rely upon in three of the six scenarios, and these varied by scenario. A comparison of the mean scores and standard deviations across participants within each scenario, and across scenarios, yields these additional observations:

– The relative cross-scenario ranking of *Competence* and *Safety* are indistinguishable, with the exception of the Home Healthcare scenario where *Safety* is one standard deviation above *Competence*. Overall, this pair is ranked as most important to participants' choice of agent in all scenarios.
– The relative cross-scenario ranking of *Predictability* and *Openness* are also indistinguishable with the sole exception of the Lost At Sea scenario where *Predictability* is one standard deviation above *Openness*.
– Small differences in ranking of agent quality categories are evident across scenarios and this may be worthy of further study.

5 General Discussion

When considering the twenty-eight specific qualities of intelligent, autonomous agents related to trust (see Table 2), the study found that the top three agent qualities cited by participants as the most important for delegation and reliance upon an autonomous agent (see Table 3) are consistent with two general qualities highlighted by previous interpersonal trust and human factors studies, i.e., (1) the ability of the machine to achieve the desired results, and (2) not causing harm.

Surprisingly, those top three qualities, consciously selected, were *not* significantly correlated with any of the actual delegation choices made by participants later in the survey when they considered specific use-case scenarios. Other lesser important qualities proved to be better predictors, and these varied by scenario (see Table 4). Our interpretation is that there may exist an influential disposition of beliefs regarding trustworthiness that are not necessarily the most salient during conscious introspection. Secondly, the context-independent responses differ from those following presentation of a challenge scenario because the participants are forced to examine their trust-related beliefs in a specific situation. These differences could prove important for future requirements and evaluation of autonomy technology; both context-independent and specific beliefs are likely to be important.

We expected *Competence* and *Predictability* to be reported by participants as the most important quality categories for trust. This was confirmed. However, only one or two of these four categories proved in any single use-case scenario to be a good predictor of choice of reliance on an

Table 9 Participant Report of Rank Order by Importance of Autonomous Agent Quality to Participants Decision to Become Reliant on an Agent[h]

Scenario	Category Rank Order				Grand Mean	*SD*
Airport Trans.	*Safety*, $\bar{x} = 3.69$	*Competence*, $\bar{x} = 3.56$	*Predictability*, $\bar{x} = 2.59$	*Openness*, $\bar{x} = 2.38$	3.05	0.67
Financial Man.	*Competence*, $\bar{x} = 3.56$	*Safety*, $\bar{x} = 3.06$	*Openness*, $\bar{x} = 2.34$	*Predictability*, $\bar{x} = 1.94$	2.73	0.73
Medical Proc.	*Safety*, $\bar{x} = 3.75$	*Competence*, $\bar{x} = 3.72$	*Predictability*, $\bar{x} = 2.50$	*Openness*, $\bar{x} = 2.25$	3.05	0.79
Home Health.	*Safety*, $\bar{x} = 3.78$	*Competence*, $\bar{x} = 3.28$	*Predictability*, $\bar{x} = 2.97$	*Openness*, $\bar{x} = 2.75$	3.20	0.45
Disaster Resp.	*Competence*, $\bar{x} = 3.56$	*Safety*, $\bar{x} = 3.41$	*Predictability*, $\bar{x} = 2.59$	*Openness*, $\bar{x} = 2.56$	3.03	0.53
Lost at Sea	*Competence*, $\bar{x} = 3.66$	*Safety*, $\bar{x} = 3.47$	*Predictability*, $\bar{x} = 2.72$	*Openness*, $\bar{x} = 2.16$	3.00	0.69

[h] Rank proportional to Mean Score $\bar{x}$ across participants (4 = Very Important; 3 = Important; 2 = Somewhat Important; 1 = Not at all Important).

intelligent, autonomous agent, and then in only in three of the six scenarios (see Table 5).

Two of the most predictive qualities were also strongly correlated with participants' overall assessment of the "ethos" of the intelligent, autonomous agents in the scenarios (per measured Source Credibility factors; see Table7). This provides further evidence for the importance of these qualities to an overall positive perception of the agent.

We conclude that self-report of the individual importance of specific trust-related qualities of intelligent, autonomous agents without context are poor predictors of a decision to rely upon such an agent when a person is confronted with a choice in a specific hypothetical use-case scenario. Other qualities appear to be elevated in importance in specific scenarios based on a combination of situational and personality factors. This points towards a need for future studies that examine trust-challenging use case scenarios using methods that achieve greater realism. In particular, such scenarios should potentiate affective responses that cannot be examined easily using surveys or unrealistic in-laboratory experiments. Specific individual personality factors were shown to influence the choice to become reliant on an autonomous agent. These include likelihood of accepting innovation, perception and acceptance of different types of risks, and factors such as *Extraversion*, *Openness*, and *Conscientiousness*.

The importance of these results for developers of intelligent, autonomous agents, perhaps embodied as robots, is likely to have a significant impact on human-robot interaction design choices. Understanding how the population of innovators and early adopters targeted by this study respond to challenging questions of trust will be critical for acceptance and deployment in critical applications of broad interest such as rescue robotics. Further investigations are required to understand the degree to which these results extend to different populations who might reasonably be expected to interact with intelligent, autonomous agents, in particular naïve users.

Technology developers should consider how the qualities of *Competence*, *Predictability*, *Safety* and *Openness* of intelligent, autonomous agents are accurately measured and portrayed in the human-robot interface. Portrayal must be done in such a manner that it correctly evokes human interpersonal trust evaluative processes and furthermore, contributes to well-calibrated trust and appropriate reliance in specific challenging application scenarios. Related follow-on research by the authors is underway using an immersive simulation of a disaster environment and interaction with a social robot to improve the methodology and further examine attributions of trustworthiness [3].

Key Points:

- We did not confirm any scenario-independent specific agent qualities that uniformly contributed to an affirmative human reliance decision. Certain qualities were important in some scenarios and not in others.
- Specific qualities of agents, and categories of those qualities, are likely to be raised or lowered in importance depending on both situational (application-specific) factors and human psychological factors. Further investigation is required to identify those qualities and provide a mechanism for understanding how their role and importance in human reliance decisions changes.
- Certain personality factors, including high scores for *Extraversion*, *Openness*, and *Conscientiousness* are very important in some situations while in others, the tolerance for certain kinds of risk is dominant when it comes to deciding to become reliant on an intelligent, autonomous agent. This suggests that certain people may more readily accept a dependency on an agent, and conversely, others are likely to be extremely resistant. Heretofore, most research on human-robot interaction has not considered the importance of individual differences.

6 Limitations

We acknowledge several limitations in our study that taken individually and as a group should provide caution regarding extrapolation of these results to other or larger populations. Despite these limitations, however, the exploratory nature of our study provides useful data and succeeds in provoking a need for further research.

Sample Population. Our sample of subject matter experts may not be representative of the larger population and this

adds uncertainty to our conclusions. The familiarity of this group with the technology of intelligent, autonomous agents may bias their attitudes, inoculate them from some degree of unconscious anthropomorphism, and increase their overall distrust. This question deserves investigation.

Agent Qualities. The twenty-eight qualities of intelligent, autonomous agents are described using words that may have different meanings to different sample populations. As such, they can only be approximations of the semantics of actual beliefs about agents.

Statistical Measures and Error. While we have applied appropriate statistical analysis to the data in our study and been as rigorous as possible with control of the data, it is possible that errors have been introduced from various sources. These include participant error, programming error and data coding error.

Survey Technical Limitations. The automation used for development and administration of the survey did not permit certain procedures that are often used to improve validity of the data collected. Specifically, the survey questions were in fixed order of presentation, identical for all subjects and not randomized. Secondly, the survey questions were not able to use a randomized mixture of positive and negative phrasing.

7 Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgements The research described in this article and preparation of this manuscript was supported by a grant from the United States Government, Air Force Office of Scientific Research (FA9550-12-0097), David J. Atkinson, Ph.D, Principal Investigator.

References

1. Ambady N, Weisbuch M (2010) Nonverbal behavior. In: Fiske ST, Gilbert DT, Gardner L (eds) Handbook of Social Psychology, vol 1, 5th edn, John Wiley & Sons, pp 464–497
2. Atkinson DJ, Clark MH (2013) Autonomous agents and human interpersonal trust: Can we engineer a human-machine social interface for trust? In: Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium, AAAI Press, SS-13-07, pp 2–7
3. Atkinson DJ, Clark MH (2014) Methodology for study of human-robot social interaction in dangerous situations. In: Proceedings of the 2nd International Conference on Human-Agent Interaction, ACM Press, pp 371–376
4. Beck HP, Dzindolet MT, Pierce LG (2002) Operators' automation usage decisions and the sources of misuse and disuse. In: Sals E (ed) Advances in Human Performance and Cognitive Engineering Research, vol 2, Emerald Group Publishing Ltd., Bingley, England, pp 37–78
5. Benet-Martinez V, John OP (1998) *Los Cinco Grandes* across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. J Pers Soc Psychol 75(3):729–750
6. Blais AR, Weber EU (2006) A domain-specific risk-taking (DOSPERT) scale for adult populations. Judgm Decis Mak 1(1):33–47
7. Carruthers P, Smith PK (eds) (1996) Theories of Theories of Mind. Cambridge University Press, Cambridge, England
8. Castelfranchi C (2000) Artificial liars: Why computers will (necessarily) deceive us and each other. Ethics Inf Technol 2(2):113–119
9. Compagni A, Mele V, Ravasi D (2015) How early implementatinos influence later adoptions of innovation: Social positioning and skill reproduction in the diffusion of robotic surgery. Acad Manage J 58(1):242–278
10. Cramer H, Evers V, Kemper N, Wielinga B (2008) Effects of autonomy, traffic conditions and driver personality traits on attitudes and trust towards in-vehicle agents. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Press, vol 3, pp 477–482
11. Cramer H, Goddijn J, Wielinga B, Evers V (2010) Effects of (in)accurate empathy and situational valence on attitudes towards robots. In: Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction, IEEE Press, pp 141–142
12. DeSteno D, Breazeal C, Frank RH, Pizarro D, Baumann J, Dickens L, Lee JJ (2012) Detecting the trustworthiness of novel partners in economic exchange. Psychol Sci 23(12):1549–1556
13. Dunn J, Schweitzer M (2005) Feeling and believing: The influence of emotion on trust. J Pers Soc Psychol 88(5):736–748
14. Falcone R, Castelfranchi C (2001) Social trust: A cognitive approach. In: Castelfranchi C, Tan YH (eds) Trust and Deception in Virtual Societies, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 55–90
15. Feltovich PJ, Bradshaw JM, Clancey WJ, Johnson M (2007) Toward an ontology of regulation: Socially-based support for coordination in human and machine joint activity. In: O'Hare GMP, Ricci A, O'Grady MJ, Dikenelli O (eds) Engineering Societies in the Agents World VII: 7th International Workshop, ESAW 2006 Dublin, Ireland, September 6–8, 2006, Revised Selected and Invited Papers, LNCS, vol 4457, Springer-Verlag, pp 175–192

16. Gabarro JJ (1978) The development of trust, influence and expectations. In: Athos AG, Gabarro JJ (eds) Interpersonal Behavior: Communication and Understanding in Relationships, Prentice-Hall, Englewood Cliffs, NJ, pp 290–303
17. Golembiewski RT, McConkie M (1975) The centrality of interpersonal trust. In: Cooper CL (ed) Theories of Group Processes, John Wiley & Sons, pp 131–185
18. Greve HR, Seidel MDL (2015) The thin red line between success and failure: Path dependence in the diffusion of innovative production technologies. Strategic Manage J 36(4):475–496
19. Groom V, Srinivasan V, Bethel C, Murphy R, Dole L, Nass C (2011) Responses to robot social roles and social role framing. In: Proceedings of the International Conference on Collaboration Technologies and Systems, IEEE Press, pp 194–203
20. Hancock PA, Billings DR, Schaefer KE, Chen JYC, de Visser E, Parasuraman R (2011) A meta-analysis of factors affecting trust in human-robot interaction. Hum Factors 53(5):517–527
21. Hurt HT, Joseph K, Cook CD (1977) Scales for the measurement of innovativeness. Hum Commun Res 4(1):58–65
22. Jacoby J, Kaplan LB (1972) The components of perceived risk. In: Proceedings of the 3rd Annual Conference of the Association for Consumer Research, Association for Consumer Research, Chicago, IL, pp 382–393
23. John OP, Naumann LP, Soto CJ (2008) Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In: John OP, Robins RW, Pervin LA (eds) Handbook of Personality: Theory and Research, 3rd edn, Guilford Press, New York, NY, pp 114–158
24. Knapp ML, Hall JA, Horgan TG (2013) Nonverbal Communication in Human Interaction, 8th edn. Wadsworth, Belmont, CA
25. Lee JD, See KA (2004) Trust in automation: Designing for appropriate reliance. Hum Factors 46(1):50–80
26. Levin DZ, Cross RL, Abrams LC (2002) Why should I trust you? Antecedents of trust in a knowledge transfer context. In: Academy of Management meetings, Denver, CO, presentation
27. Madsen M, Gregor S (2000) Measuring human-computer trust. In: Proceedings of the 11th Australasian Conference on Information Systems, pp 6–8
28. Marble J, Bruemmer D, Few D, Dudenhoeffer D (2004) Evaluation of supervisory vs. peer-peer interaction with human-robot teams. In: Proceedings of the 37th Hawaii International Conference on System Sciences, IEEE Press, vol 5, p 50130b
29. Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. Acad Manage Rev 20(3):709–734
30. McAllister DJ (1995) Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. Acad Manage J 38(1):24–59
31. McCroskey JC, Teven JJ (1999) Goodwill: A reexamination of the construct and its measurement. Commun Monogr 66(1):90–103
32. McKnight DH, Chervany NL (2000) What is trust? A conceptual analysis and an interdisciplinary model. In: Chung MH (ed) Proceedings of the Americas Conference on Information Systems, pp 827–833
33. McKnight DH, Chervany NL (2001) Trust and distrust definitions: One bite at a time. In: Falcone R, Singh M, Yao-Hua T (eds) Trust in Cyber-Societies: Integrating the Human and Artificial Perspectives, LNCS, vol 2246, Springer-Verlag, pp 27–54
34. Muir BM (1994) Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics 37(11):1905–1922
35. Nass C, Fogg BJ, Moon Y (1996) Can computers be teammates? Int J Hum-Comput St 45(6):669–678
36. Oleson KE, Billings DR, Kocsis V, Chen JYC, Hancock PA (2011) Antecedents of trust in human-robot collaborations. In: 2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, IEEE Press, pp 175–178
37. Parasuraman R, Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. Hum Factors 39(2):230–253
38. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1(4):515–126
39. Rammstedt B, John OP (2007) Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. J Res Pers 41(1):203–212
40. Rogerson M, Gottlieb M, Handelsman M, Knapp S, Younggren J (2011) Nonrational processes in ethical decision making. Am Psychol 66(7):614–623
41. Roselius T (1971) Consumer ranking of risk reduction methods. J Marketing 35(1):56–61
42. Schaefer K, Billings D, Hancock P (2012) Robots vs. machines: Identifying user perceptions and classifications. In: 2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, pp 168–171
43. Schoorman FD, Mayer R, Davis J (2007) An integrative model of organizational trust: Past, present, and future. Acad Manage Rev 32(2):344–354
44. Stokes C, Lyons J, Littlejohn K, Natarian J, Case E, Speranza N (2010) Accounting for the human in cyberspace: Effects of mood on trust in automation. In: Proceedings of the 2010 International Symposium on Colaborative Technologies and Systems, IEEE Press, pp 180–187

45. USAF (2010) Technology horizons: A vision for Air Force science & technology during 2010–2030. Tech. Rep. AF/ST-TR-10-01-PR, United States Air Force, Office of Chief Scientist, Washington, DC
46. Wagner A (2009) The role of trust and relationships in human-robot social interaction. PhD thesis, Georgia Institute of Technology, Atlanta, GA
47. Weber EU, Blais AR, Betz NE (2002) A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. J Behav Decis Mak 15(4):263–290
48. Weber JM, Malhotra D, Murnighan JK (2004) Normal acts of irrational trust: Motivated attributions and the trust development process. Res Organ Behav 26:75–101

Figure 3 image file
Click here to download high resolution image

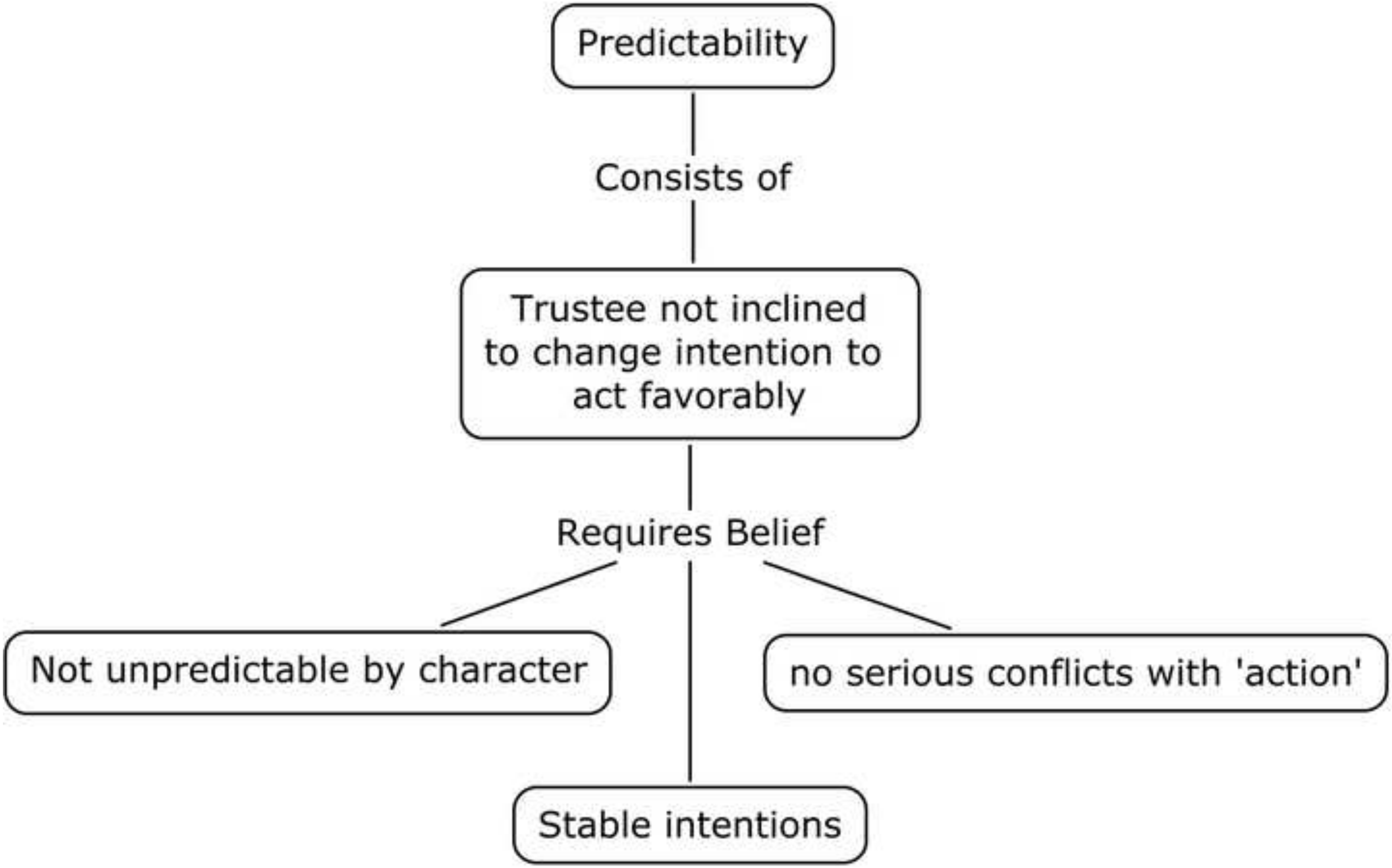


Figure 4 image file
Click here to download high resolution image

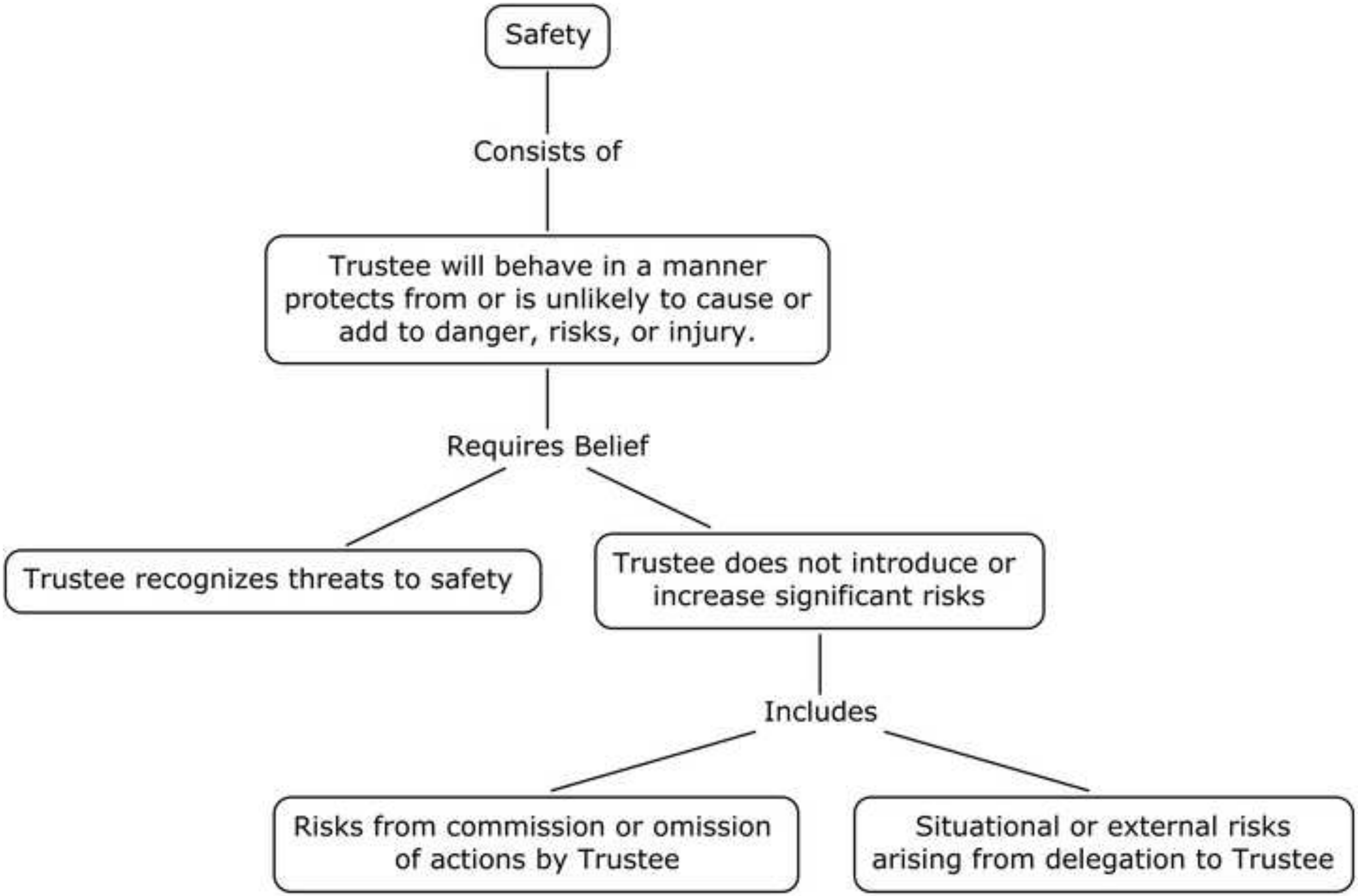


Figure 5 image file
Click here to download high resolution image

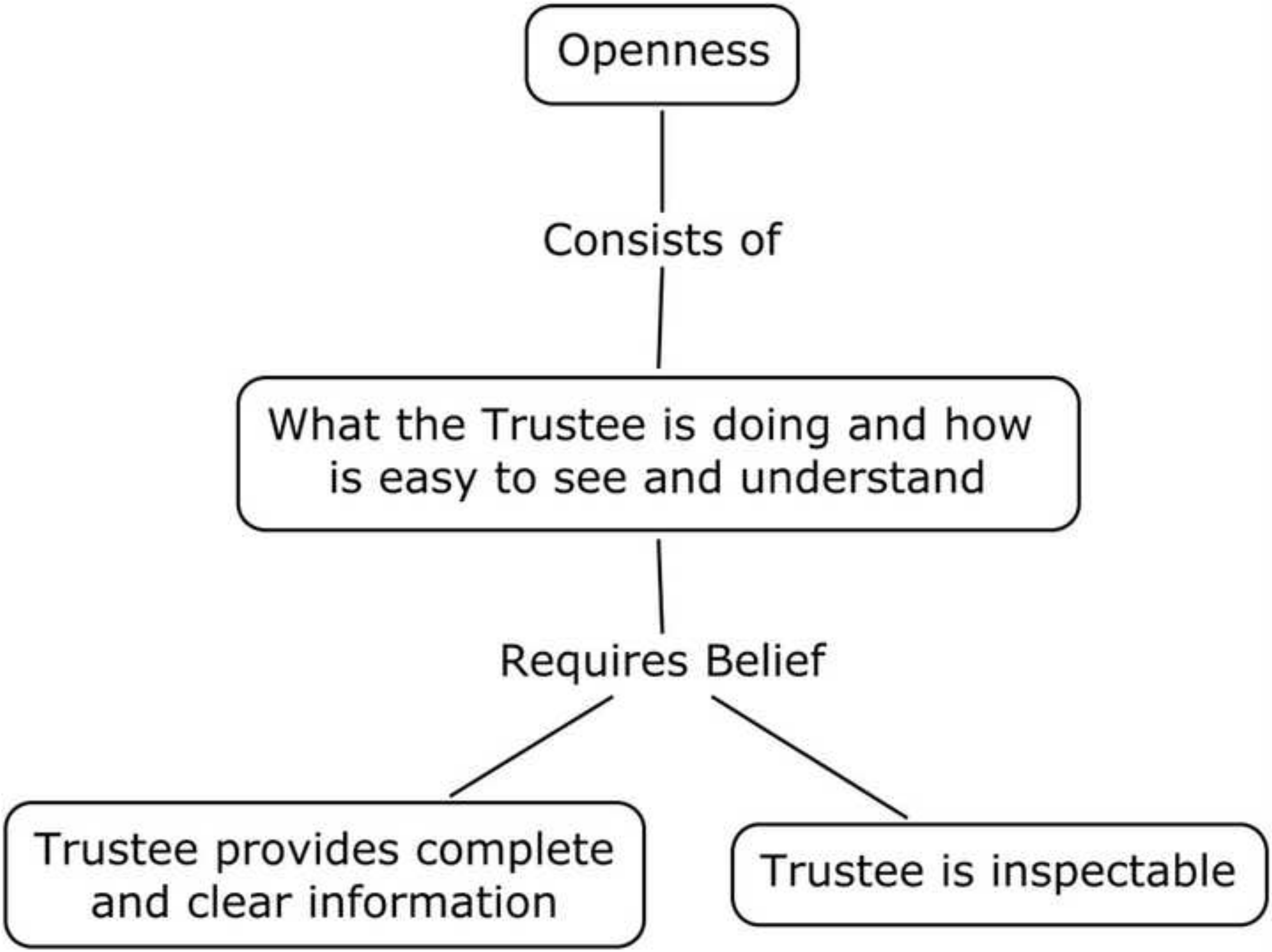


Figure 1 image file
Click here to download high resolution image

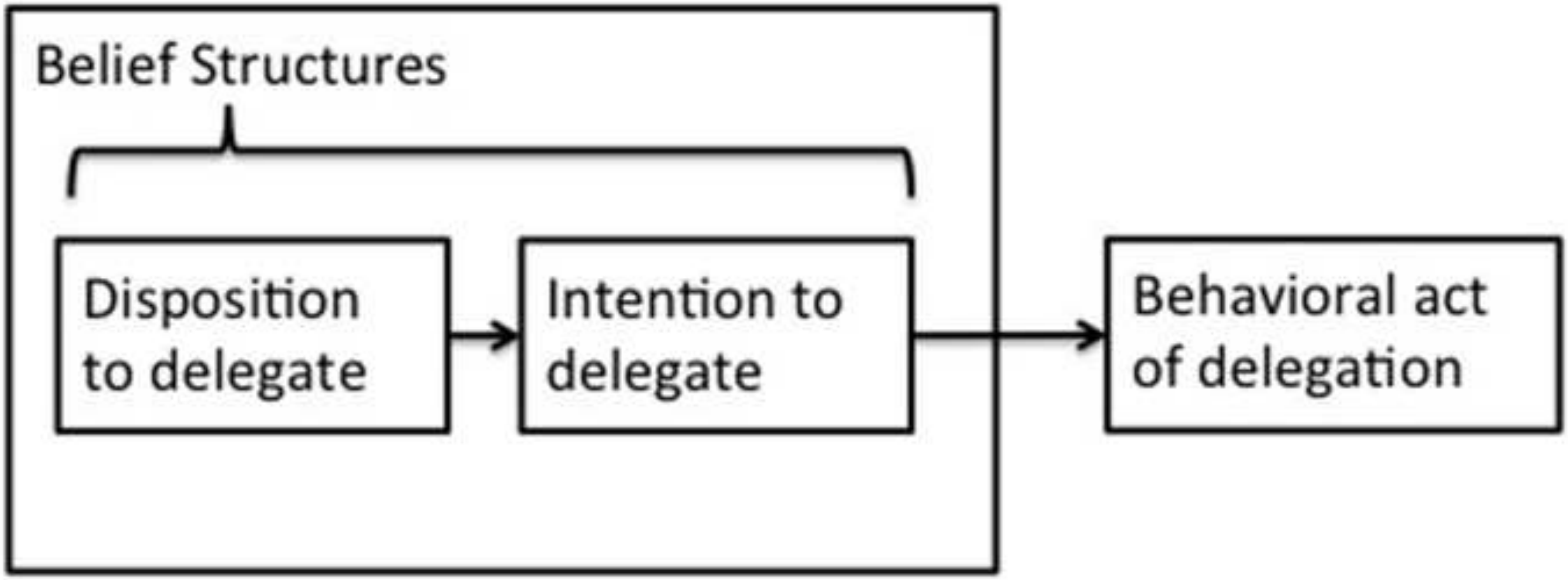


Figure 2 image file
Click here to download high resolution image

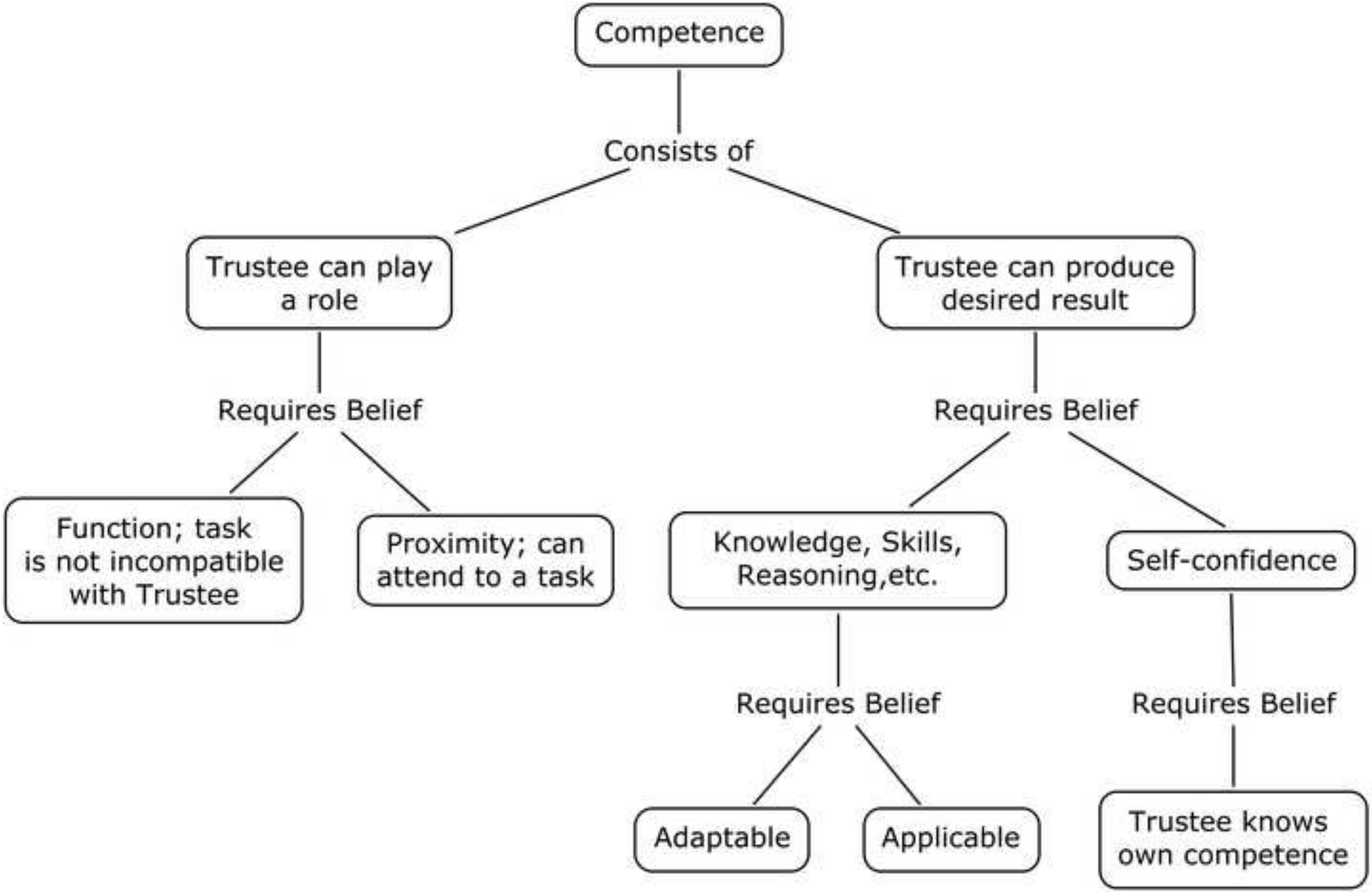


Figure 2 image file
Click here to download high resolution image

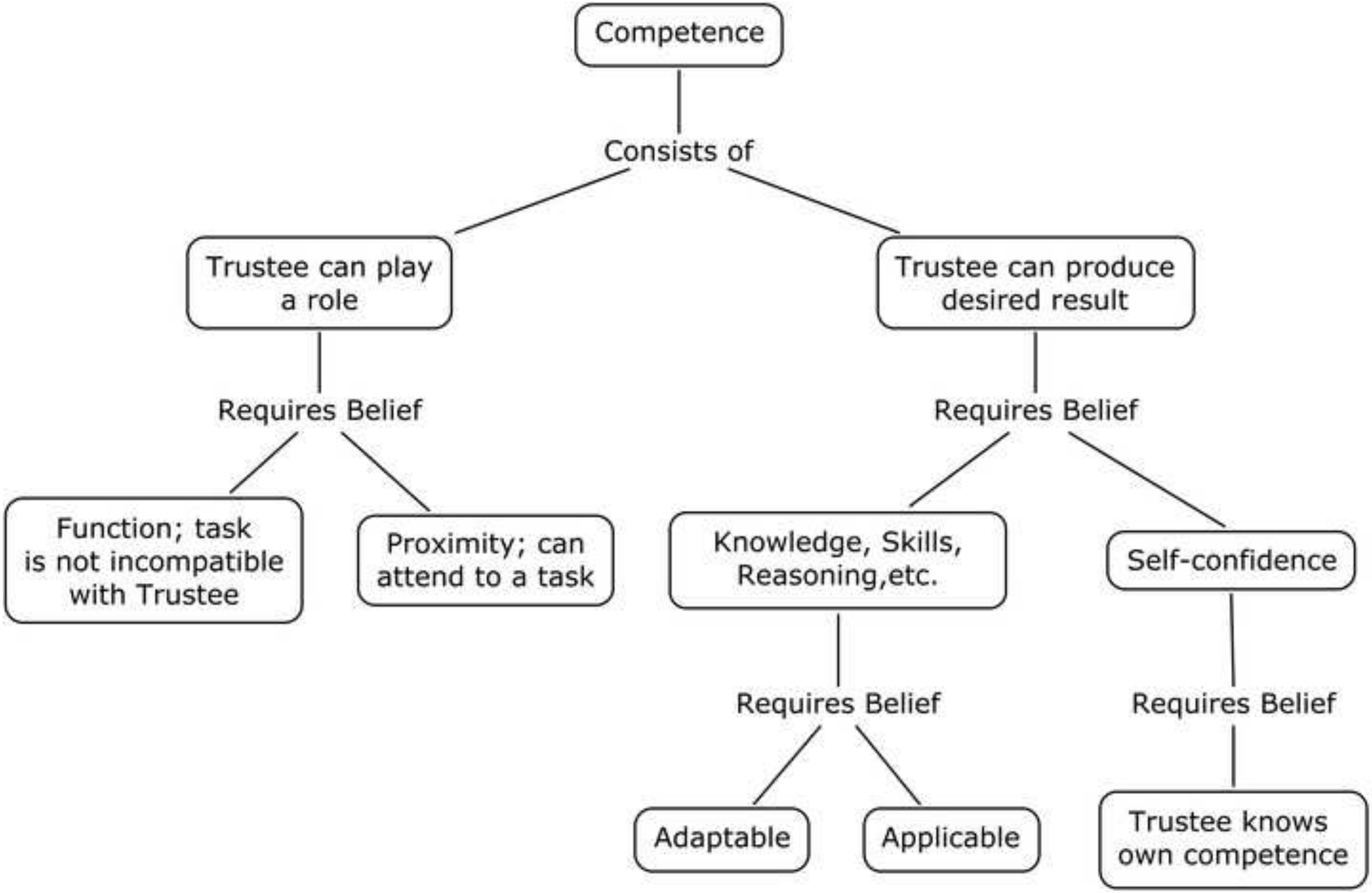


Figure 5 image file
Click here to download high resolution image

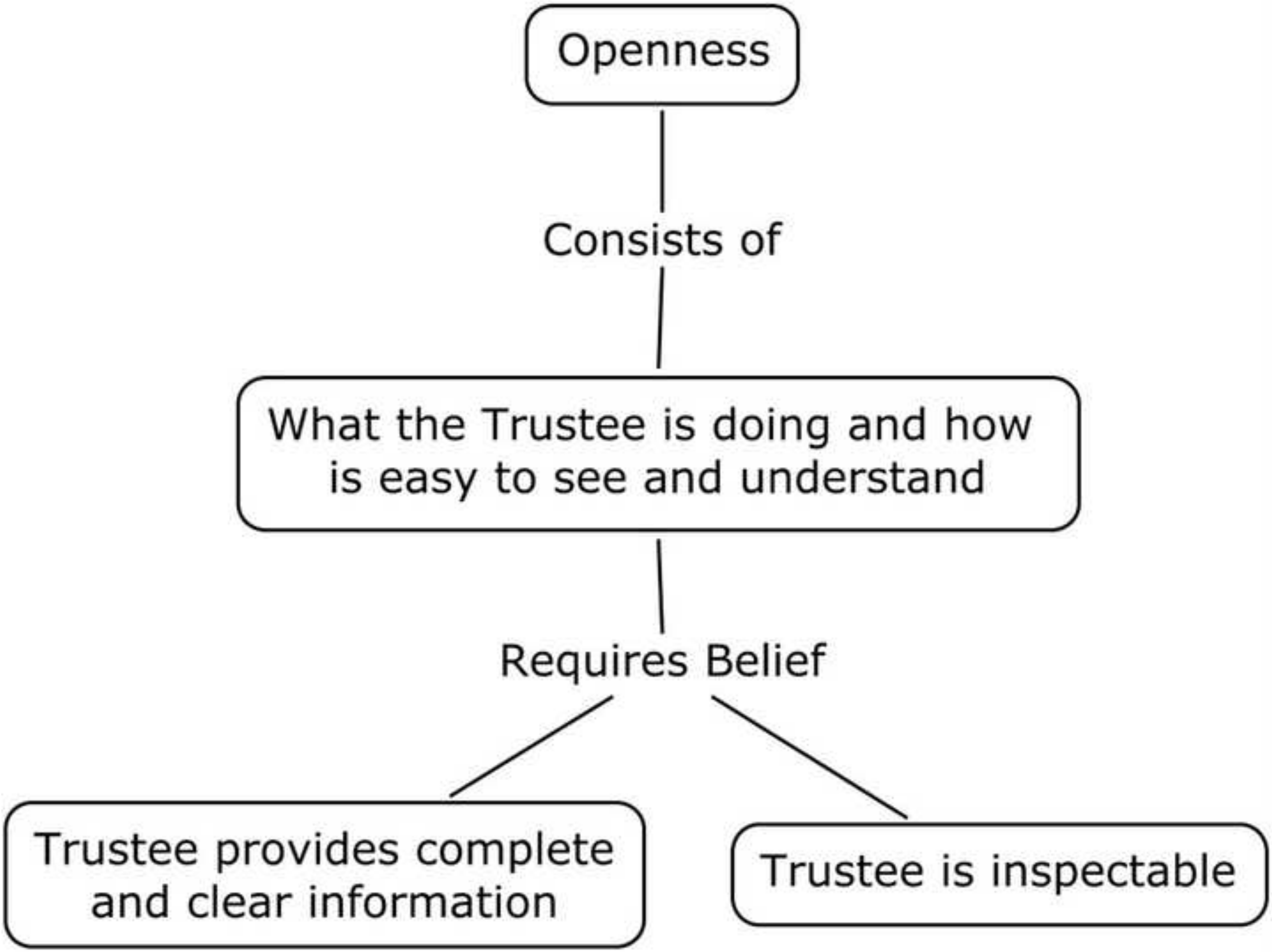


Figure 4 image file
Click here to download high resolution image

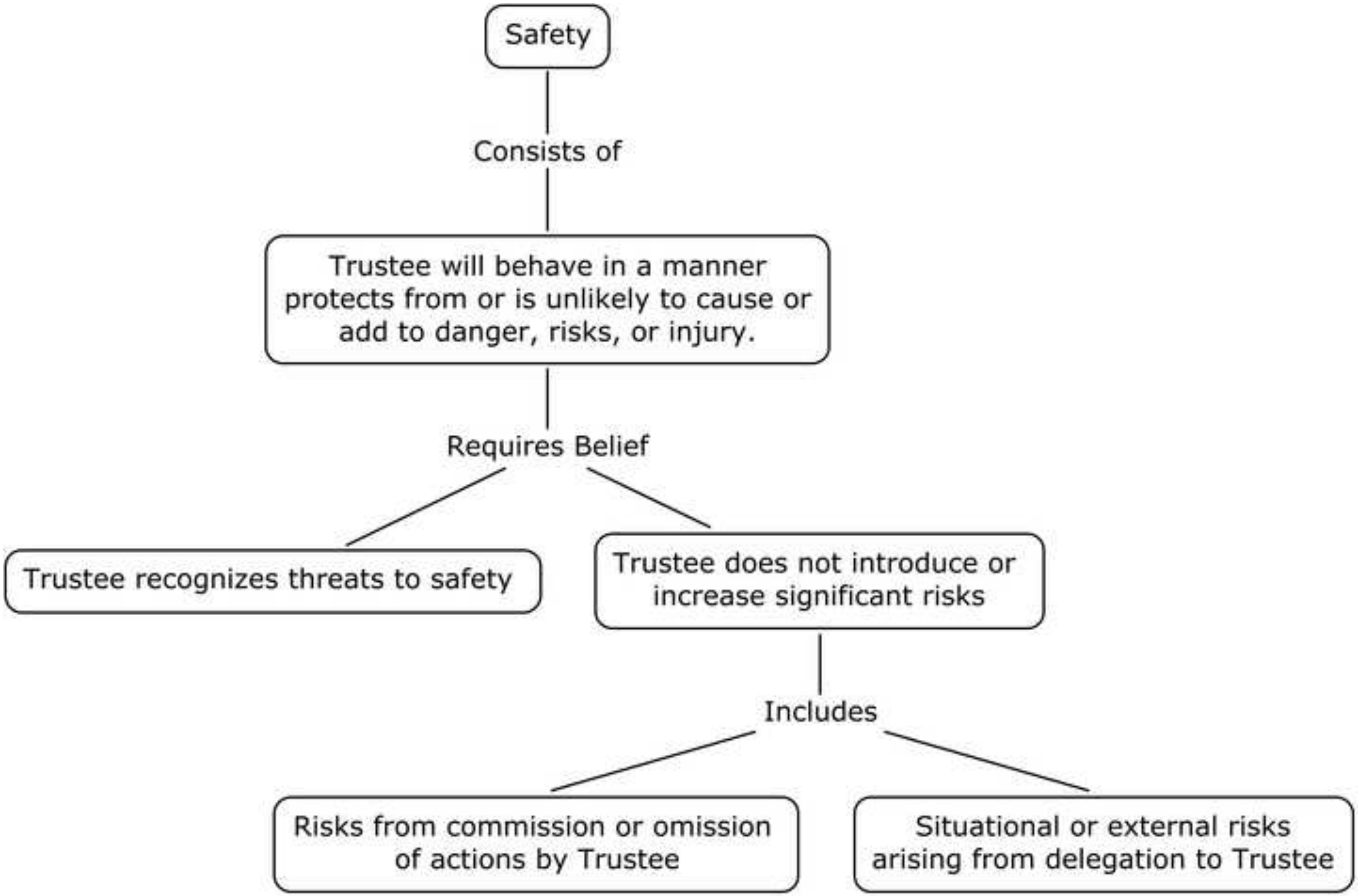


Figure 6 image file
Click here to download high resolution image

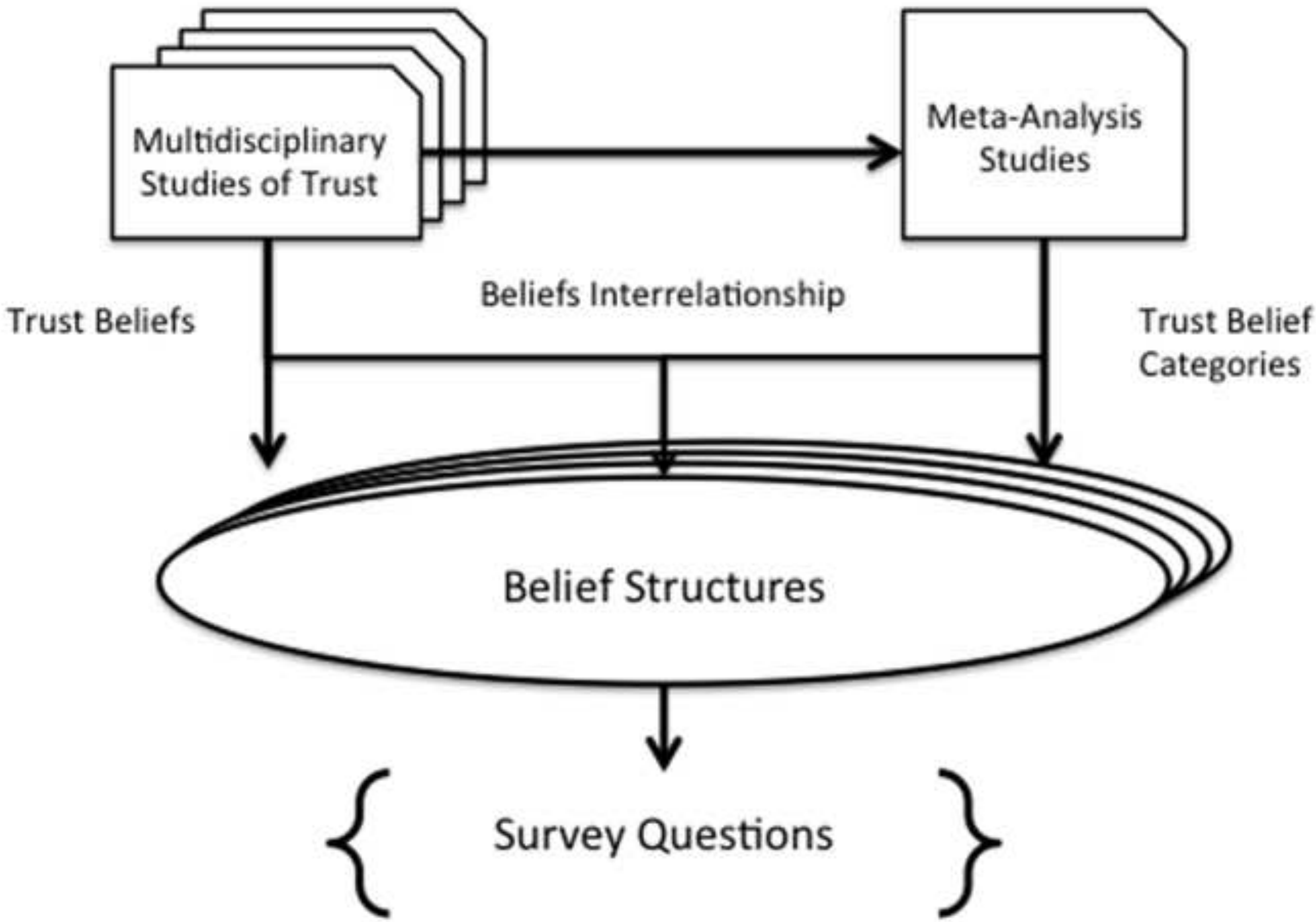


Figure 3 image file
Click here to download high resolution image

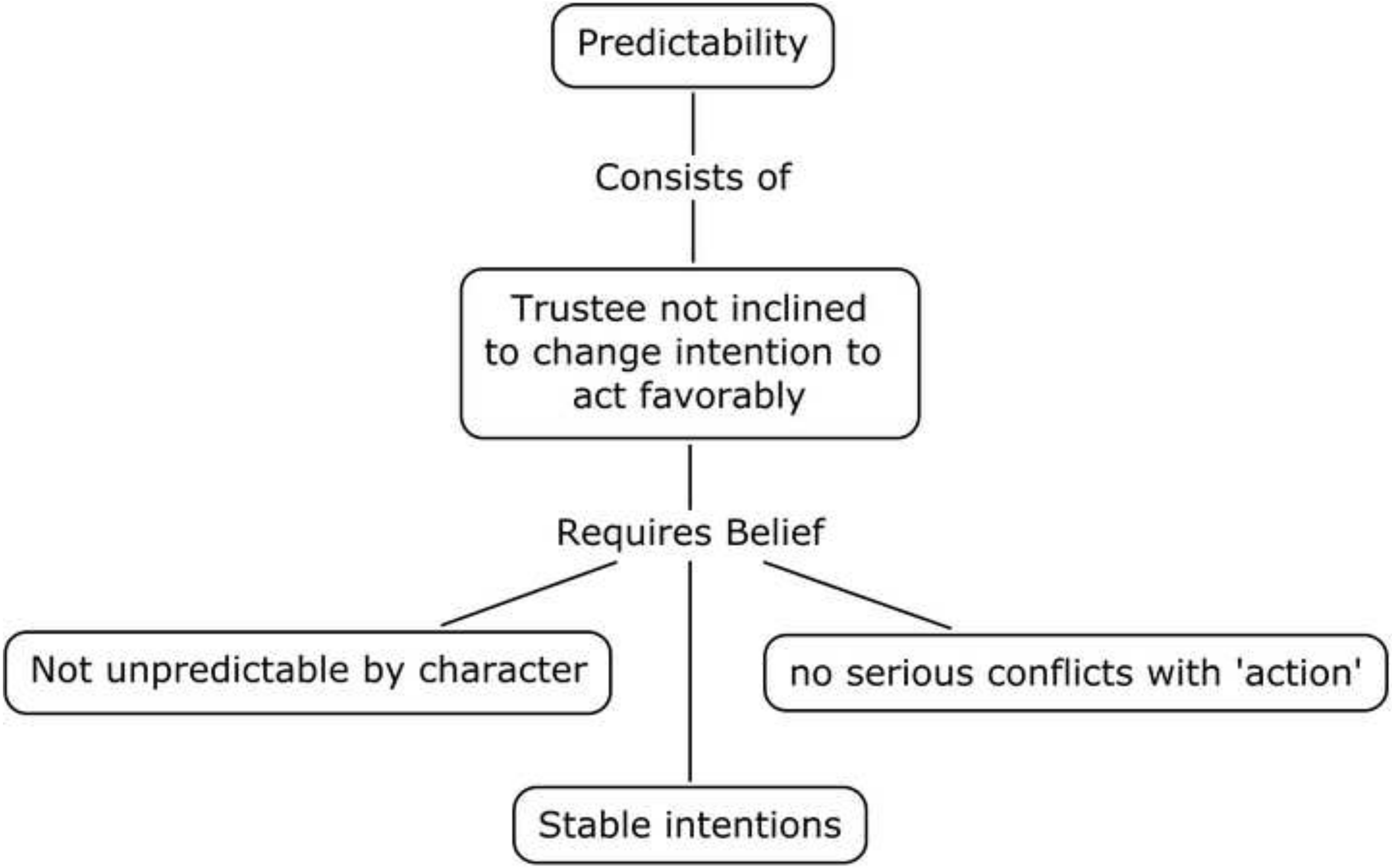


Figure 1 image file
Click here to download high resolution image

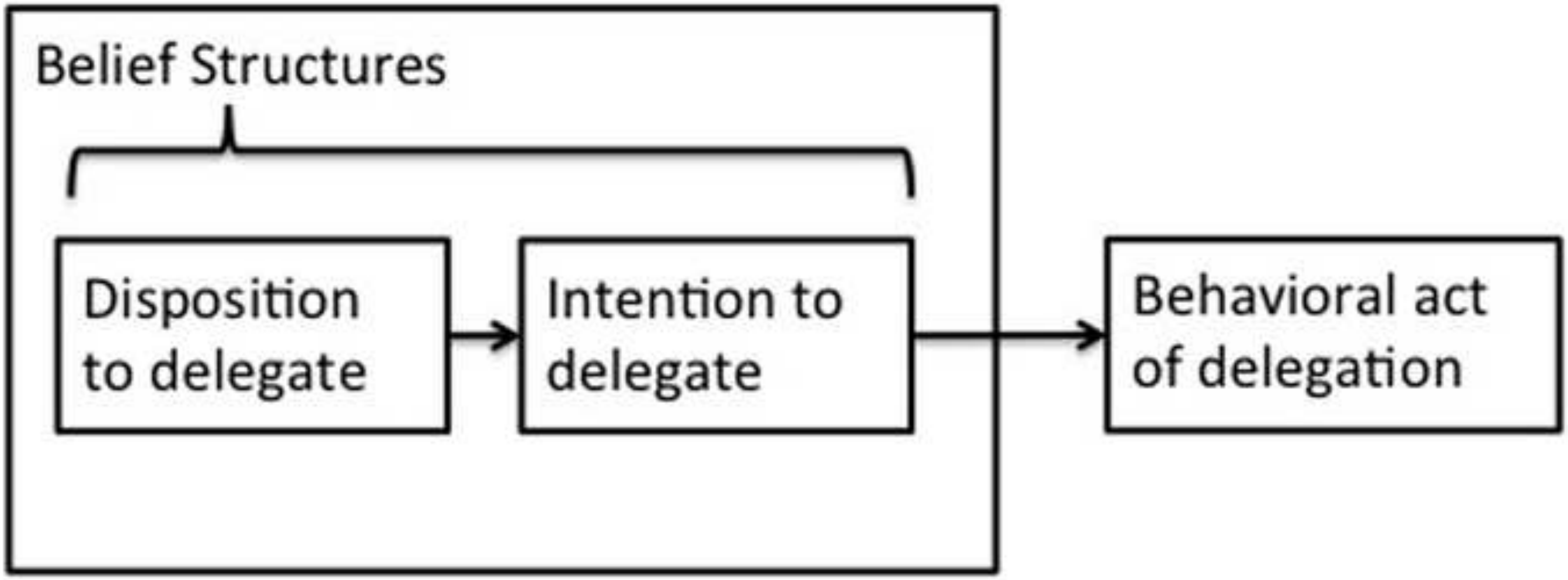
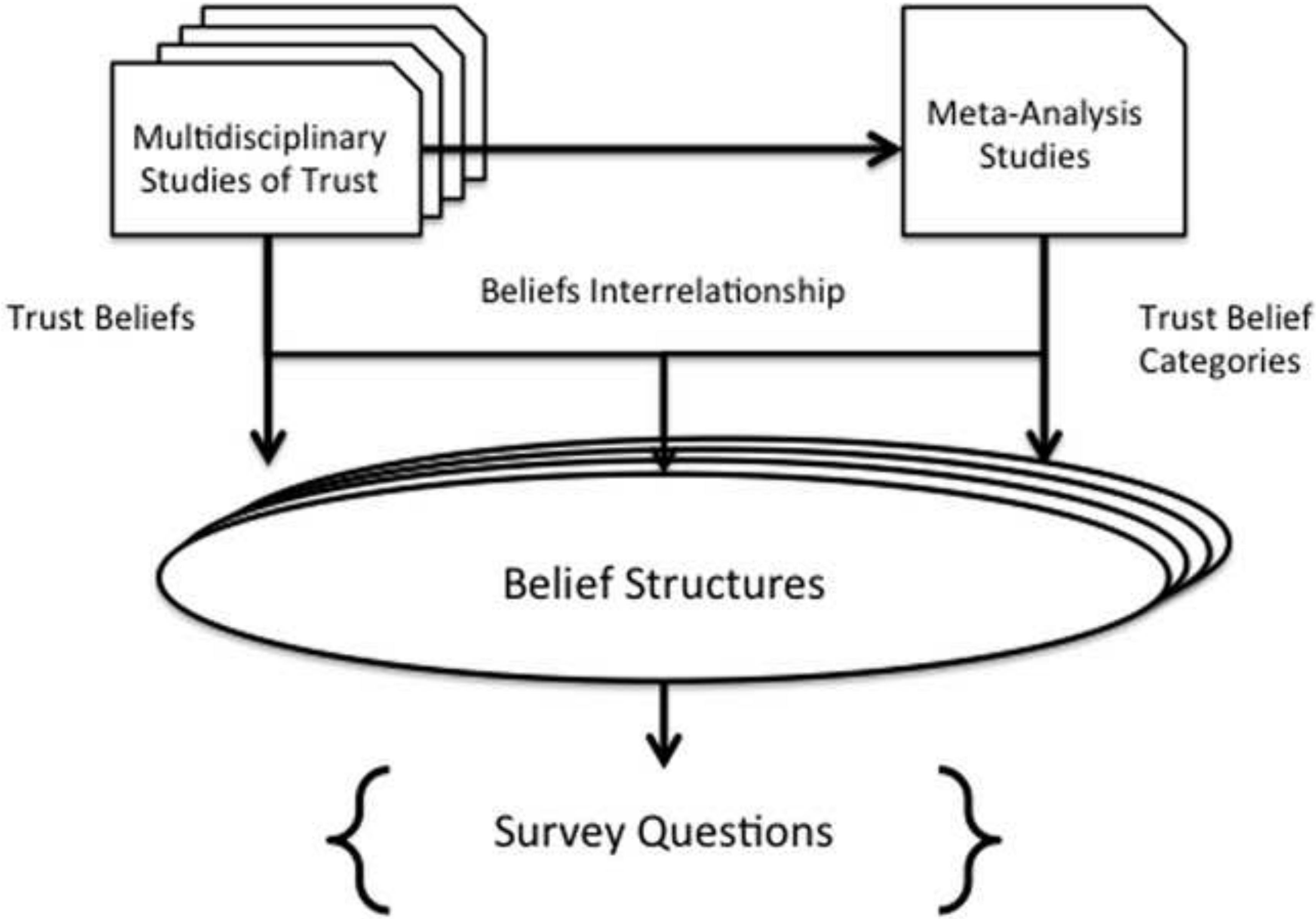


Figure 6 image file
Click here to download high resolution image



Author Preprint Copy, To Appear:
Atkinson, David J. 2015. Emerging Cyber-Security Issues of Autonomy and the Psychopathology of Intelligent Machines. In *Foundations of Autonomy and Its (Cyber) Threats, Papers from the 2015 AAAI Spring Symposium.* Technical Report No.. SS-15-01. Menlo Park, CA: AAAI Press.

Emerging Cyber-Security Issues of Autonomy and the Psychopathology of Intelligent Machines

David J. Atkinson
Institute for Human and Machine Cognition, 15 SE Osceola Avenue, Ocala, FL 34471
datkinson@ihmc.us

Abstract

The central thesis of this paper is that the technology of intelligent, autonomous machines gives rise to novel fault modes that are not seen in other types of automation. As a consequence, autonomous systems provide new vectors for cyber-attack with the potential consequence of subversion, degraded behavior or outright failure of the autonomous system. While we can only pursue the analogy so far, maladaptive behavior and the other symptoms of these fault modes in some cases may resemble those found in humans. The term "psychopathology" is applied to fault modes of the human mind, but as yet we have no equivalent area of study for intelligent, autonomous machines. This area requires further study in order to document and explain the symptoms of unique faults in intelligent systems, whether they occur in nominal conditions or as a result of an outside, purposeful attack. By analyzing algorithms, architectures and what can go wrong with autonomous machines, we may a) gain insight into mechanisms of intelligence; b) learn how to design out, work around or otherwise mitigate these new failure modes; c) identify potential new cyber-security risks; d) increase the trustworthiness of machine intelligence. Vigilance and attention management mechanisms are identified as specific areas of risk.

Introduction

Psychopathology is the study of mental illness, mental distress, and abnormal or maladaptive behavior. It is the study of *fault modes* of the human mind. As yet, we have no equivalent area of study for intelligent, autonomous machines. Software engineering techniques for reliable systems are applicable (as they are to all complex software artifacts), but insufficient. The topic of this paper is the proposition that the technology of intelligent, autonomous systems gives rise to novel fault modes that are not seen in other types of automation. These fault modes arise from the *nature of the algorithms* and how they perform in real-world situations (including human interaction) with uncertain data. As a consequence, autonomous systems may provide new vectors for cyber-attack that could lead to subversion, degraded behavior or outright system failure.

This paper arose from a bit of fun the author was having by examining examples of "robots run amok" in popular literature and media. HAL 9000, of the movie "2001: A Space Odyssey" is a canonical example. These cases are often described in anthropomorphic terms related to human psychopathology, and this became the genesis of the idea for a psychopathology of intelligent machines. Although the analogy will stretch only so far, the search for intelligent machine near-equivalents of certain human mental disorders has already yielded a few insights that are described herein. The over-riding question is whether something like the behavior of these fictional malevolent machines could actually occur. In many cases, the answer is "probably not" but in a few, the answer is "probably yes." If so, can we identify plausible mechanisms that explain the nature of the amok machines' failures, given present artificial intelligence technology and what we can reasonably project on the horizon?

This possibility suggests that there are fault modes for autonomous systems that remain unexplored and their implications unknown. The purpose of this paper is to raise that question explicitly, and to do so in the context of fault modes as potential vulnerabilities to attack, exploitation and subversion.

We stand to gain certain benefits by analyzing the unique fault modes of autonomous systems. Such studies might provide insight into aspects of machine intelligence just as studies of human mental disorders have historically provided insight into the functioning of the brain. With the human mind, psychologists seek explanations for mental disorders from biological sources (relatively rare), innate biases, and faulty inference. Such failures of the human

mind are often based in experience and learned behavior, including interpersonal communication and relationships with social and group effects. These are well documented. In contrast, with autonomous systems we must seek explanations for anomalous, maladaptive behavior in hardware (probably rare), software algorithms, logic, knowledge and situational uncertainty. Also guided by the study of human mental disorders, we should look for sources of machine intelligence fault modes in experience (episodic memory) and machine learning, including human-machine interaction and other aspects of social and affective computing.

Some of these autonomous system faults may occur in the course of day-to-day nominal operations and be easily "cured." Of greater concern, it is possible that some psychopathologies of machine intelligence could be *induced* in a new form of cyber-attack, thereby creating new risks with potentially very serious consequences. We have the opportunity, now, to focus research on how to design out, work around or otherwise mitigate the failure modes we discover. It is best if this is accomplished sooner rather than later due to the potential adverse consequences. Ultimately, the real payoff for AI research and development of autonomy applications is the opportunity to increase the trustworthiness of machine intelligence. Today, this is cited as a chief obstacle to greater deployment of autonomous systems (Dahm 2010).

The sections below provide essential background and an initial analysis of the symptoms and sources of selected example fault modes of autonomous, intelligent systems. In each case, we examine these fault modes with respect to vulnerability to cyber-attack. In the conclusion section, we discuss directions for future research and parameters of the required studies.

Essential Background

The technology of autonomous systems extends beyond conventional automation and solves application problems using materially different algorithms and software system architectures. This technology is a result of multidisciplinary research primarily in the fields of artificial intelligence and robotics, but drawing on many other disciplines as well, including psychology, biology, mathematics and others. Research on autonomous systems spans multiple areas, including (but not limited to) algorithms, computing theory, computing hardware and software, system architectures, sensing and perception, learning, and the acquisition and use of large stores of highly interconnected and structured, heterogeneous information.

The key benefit realized from autonomy technology is the ability of an autonomous system to explore the possibilities for action and decide "what to do next" with little or no human involvement, and to do so in unstructured situations which may possess significant uncertainty. This process is, in practice, indeterminate in that we cannot foresee all possible relevant information (i.e., features and their relationship to one another) that could be a factor in pattern-directed decision-making.

The autonomous ability to decide on next-steps is the core of what enables many valuable applications. "What to do next" may include a wide variety of actions, such as: a step in problem solving, a change in attention, the creation or pursuit of a goal, and many other activities both internal to the operation of the system as well as actions in the real world (especially in the case of embedded or cyber-physical systems). Ill-informed efforts to "envelope" or otherwise externally constrain the behavior of autonomous systems are sacrificing the most important strength of the technology – to perform in ways we cannot *a priori* anticipate.

However, while the technology delivers new capabilities to perform work in a wide variety of under-specified and dynamic situations, it is also extremely complex to the point where conventional software systems test and evaluation methods are no longer sufficient to establish nor maintain confidence in autonomous systems. It is system *complexity*, arising from specific component technologies of autonomy (individually and collectively), that creates the prospect of new cyber-security risks.

Of special importance is *computational complexity*: a measure of the resources required by a given algorithm to reach a result. Computational complexity is measured in time (e.g., wall clock time) and space (e.g., memory storage), and there are multiple other important attributes as well. The decision by an autonomous system of "what to do next" is the result of an algorithm that can be viewed, abstractly, as maximizing a utility function. These algorithms, intrinsic to autonomous systems, are typically of very high computational complexity; that is, they may require *exponential* amounts of time and/or space.

Strict utility-based decision-making processes are recognized to be impossible in non-trivial domains (for people as well as machines). This is a result of the potentially infinite courses of action available, and the consequent inability to exhaustively analyze all of the near infinite number of possible system states; the inability to obtain and store all potentially relevant facts; and the intrinsically uncertain relationship between chosen actions and consequences when the environment has complex dynamics including other actors (Brundage 2014).

Consequently, as a rule, the process of decision-making by an autonomous system is intrinsically limited by the available information, computational resources, and the finite amount of time available to reach a conclusion. This is referred to as "bounded rationality" (Simon 1958) and

serves as a bedrock principle for research in artificial intelligence (AI). The result is that we can only hope to *approximate* optimal decision-making and behavior in an intelligent, autonomous system: "Satisficing" is acting in a way that leads to satisfactory and sufficient ("good enough") outcomes.

We conjecture that this heuristic, *algorithmic struggle for computational resources* with limited time and information is a principal source of novel fault modes that arise in autonomous, intelligent systems.

Fault Modes

What could possibly go wrong? That is the question asked by every researcher, developer, decision-maker, and user of an intelligent system. There exists the familiar panoply of software and system faults shared by all complex computational systems. Those are not our focus here. Our interest is in what *new* types of faults might exist *by virtue of the nature of the algorithms* in intelligent systems, or their application in certain circumstances, or as a result of *malicious manipulation*. Do such fault modes exist?

The purpose of this section is to stimulate thought, discussion, and ideally, to convince you that the answer is likely to be "yes." The existence of these fault modes arises directly from the limitations imposed on autonomy technology by computational complexity, as discussed in the previous section. Such faults are today typically conceptualized in terms of constraints on algorithms rather than cyber-security vulnerabilities; this paper aims to raise awareness of that gap in our understanding.

The systems test and evaluation community has recognized that something is really different about autonomous systems, specifically, the *near infinite number of potential system states* in an intelligent, autonomous system renders much of existing test and evaluation methodology insufficient (or at worse, ineffective) for producing high confidence assertions of performance and reliability (Dahm 2010). The ideas presented here ideally ought to lead to enhanced test and evaluation processes, but we leave that to be discussed elsewhere.

In the search for novel fault modes, we are guided by our (admittedly imperfect) analogy to human psychopathology and certain philosophical considerations. If the computational mechanisms of intelligence are independent of the physical medium that supports such computations, then what is true of one type of intelligent system may also be true of another type. This is implied by the philosophical formulation of machine-state functionalism (Putnam 1979) upon which much of artificial intelligence *and* cognitive science research is predicated.

The subsections below describe potential fault modes that may arise in an example set of functional areas common to many intelligent, autonomous systems. In each case, we would like to understand the symptomology of faults and the underlying causes. Only then can we investigate vulnerabilities, methods of detection, isolation and repair. Without presenting tutorial information best found elsewhere, we consider potential fault modes arising in the processes of:

1. Goals and Goal Generation
2. Inference and Reasoning
3. Planning and Execution Control
4. Knowledge and Belief
5. Learning

Goals and Goal Generation

Goals are the primal initiator of behavior in a *deliberative* autonomous system (in contrast to a reactive autonomous system, for example, one based on a subsumption architecture (Brooks 1988) which is driven more directly by sensory data; many autonomous systems are hybrids of deliberative and reactive components. In deliberative systems, a goal state may be completely specified, only partially specified, or may be in the form of a general preference or constraint model with "goodness" evaluated according to certain formulae. A wide variety of preference/constraint models exist, some applicable only to deterministic domains and others to probabilistic domains or where preferences must be explicitly elicted (Gelain et. al. 2009; Dalla Pozza 2011).

Some examples of candidate psychopathological fault modes related to goals that are shared with people, but not other non-intelligent machines, are: Disorders of Attention, Goal Conflict, Indifference, and Self-Motivated Behavior. We examine each of these in turn.

Disorders of Attention. The pursuit of goals, including goal generation, goal selection, and deliberative planning, all require allocation of system resources. In each of these functions, decisions are made about how to use computational resources. These decision-making processes can be viewed fundamentally as *attention management mechanisms* (Helgason, Nivel and Thorisson 2012).

Goal generation functions (triggered by external or internal information) are fundamentally *vigilance mechanisms* because they can divert attention. Diverting attention diverts the management of scarce system resources. In most cases, this is appropriate and exactly what the designers of intelligent systems intend (Coddington 2007; Hawes 2011).

With respect to cyber-security, however, this suggests that attacking vigilance mechanisms has the potential to divert attention and resources away from what an autonomous system "ought" to concentrating upon. Misappropriation or diversion of scarce computing

resources is a potential critical vulnerability of autonomous systems that may appear as a consequence of other types of faults.

Goal Conflict. The resolution of conflicting requirements for achieving different goals is a fundamental component of all AI planning and scheduling algorithms. There are many such planning algorithms, and equally many ways to resolve goal conflicts. It is important to remember that *the guaranteed detection of goal conflicts during the planning process is computationally intractable*. Heuristic methods are required in order to focus attention on likely sources of goal conflict (Luria 1987). These heuristics are also *attention management mechanisms.* Luria (op. cit.) provides a brief taxonomy of goal conflicts. Drawing from that taxonomy provides a good start towards identifying goal conflict-related fault modes (see Table 1 for examples). These modes are each potential vectors for cyber-attack by an adversary with the capability to artificially induce the conditions that enable a type of goal conflict.

TYPE OF CONFLICT	DESCRIPTION
Compromised Intent	Conflict between explicit goal and default policy or implicit intent.
Violated Defaults	Unverified knowledge of the values of default preconditions.
Unintended Effects	Plan used in a novel situation with un-modeled direct interactions.
Expressed Conflict	Human agent asserts that a conflict exists, with or without explanation.
Effects Cascade	Effects of plan execution result in an unrelated conflict (side effect), e.g., due to insufficient causal model fidelity, inference horizon, etc. If the effects are non-linear, a cascade is possible.

Table 1: Example Types of Goal Conflicts.

Consider just one of the many sources of goal conflicts that are known: *Compromised Intent*. This type of goal conflict occurs when achievement of a goal conflicts with default policy or intent. It may occur because (1) a causal interaction is not modeled, or; (2) an inference chain is too long to find the conflict (as in a search with a bounded horizon), or; (3) unknown, explicit or implicit priorities, or other conditions that enable the relaxation of constraints. I would be greatly surprised if a reader familiar with AI planning systems has not seen this type of conflict.

There is another reason it seems familiar. Return to our (fictional) example in the introduction, HAL 9000, of the movie "2001: A Space Odyssey." Recall that the super-secret, highest priority mission goal given to HAL is to investigate the monolith at Jupiter. This explicit goal comes into conflict, later in the mission, with the default policy of protecting the lives of the crew. This is just one of many types of potential goal conflicts that may not be detected before actual execution of a plan. Skipping over the drama of the movie, we discover that HAL chooses to resolve this goal conflict by killing the crew. The hypothetical mechanism is relatively easy to discern: a relaxation of a constraining default policy (crew safety) in order to achieve a high priority goal (investigate the monolith). The constraint relaxation is enabled by HAL's *certainty* that he can complete the secret mission without the aid of the crew (this is also a failure of *ethical reasoning*, discussed later). In humans, unresolved goal conflict is a source of significant mental distress (Mansell 2005). Similarly, resolution of goal conflict is often (but not always) an imperative in autonomous systems.

Indifference. This type of fault is a milder form of goal conflict that can result from an intelligent system concluding that (1) a human-provided goal has insufficient priority relative to other goals, or (2) the system itself does not possess the competence to achieve a goal, or (3) the goal is irrelevant. The consequence of goal conflict resolution is the human-provided goal is dropped and no action occurs to achieve the goal. In human terms, this condition is described as *apathy*.

Self-Motivated Behavior. Autonomy necessarily implies a degree of choice of actions, whether they originate from externally provided goals or goals internally generated. In the latter case, the addition of autonomic processes to a system (e.g., for health maintenance, energy management and so forth) can result in goals that conflict with on-going activities. The examples we see today, such as a robot vacuum cleaner stopping to recharge itself, are expected behaviors and not of interest. However, as intelligent robots are deployed into dangerous situations, such as urban rescue or a battlefield, their autonomic functions are likely to expand to include *self-preservation* as a default autonomic function. Consider the possibility that an undesirable machine behavior (perhaps as a result of another fault and/or subversion) results in a shutdown command. Due to a conflict with the internally generated goal of self-preservation, the directive to shutdown could be ignored in certain situation-driven conflict resolutions.

It is important to note that both *Indifference* and *Self-Motivated Behavior* may have the *appearance* of self-awareness. Yet, the information and goal conflict resolution processes are localized within the intelligent system. The appearance of self-awareness is an *emergent psychological effect* (Lewis et. al. 2011); actual self-awareness is not required.

Inference and Reasoning

There are several important sources of potential faults in the area of inference and reasoning that require further

study. Some are familiar to many of us from long coding and test sessions with intelligent systems, others are speculative possibilities that may arise in future systems.

Invalid Logic. Often termed "fallacies of inference", there are many forms of invalid logic that humans demonstrate. As yet, intelligent systems only suffer from a few. One of these cases is when "true" data results in a false answer as a result of a failure of inductive reasoning; for example, when an intelligent system is near the edges of its competence. As a consequence, insufficient previous experience (e.g., manifested as an incorrect probability distribution) results in over-weighted confidence for derived conclusions. This leads to the possibility that new data becomes marginalized or discarded rather than serving in a corrective function. This is an example of the classic "over-generalization" problem in machine learning, where important features that discriminate situations are ignored.

The Fallacy Fallacy. This fault mode is complementary to *Invalid Logic*. Knowledge bases are inherently incomplete, likely to contain errors, and subject to many other limitations. One potential consequence is that a conclusion is dismissed because the logic used to derive the conclusion is faulty or incomplete, i.e., there is no inference chain to the conclusion that can be constructed from the knowledge and data given (or as a result of a bounded search horizon, as discussed earlier). If the argument contains a fallacy, i.e., invalid logic, then *it is the argument that must be dismissed*. The failure to construct an inference chain does not prove that the *conclusion* is incorrect, only that it cannot be proved with what is known. The conclusion may in fact be correct. Few, if any, extent intelligent systems respect this distinction; it is a defect of reasoning that unfortunately shared by many people as well.

Solipsism. One of the dangers of the AI craft of applied epistemology arises from the quest to manage uncertainty. This has the potential to result in a sort of *logical minimalism* where sense data is subject to extreme skepticism and as a result, internally derived inferences may accrue more confidence than those based on empirical observations. In a sense, this is the robot equivalent of the human psychopathological condition of *detachment from reality*. The danger arises when solipsism undercuts externally imposed policy-guided constraints on behavior by authority.

Planning and Execution Control

There are a great many faults that can arise during the planning and execution control processes, including many of those related to goals as we have discussed above. Planning is essentially a search problem with surprising complexity that often requires exponential computation, i.e., is NP-hard (Chapman 1987; Hendler et. al. 1990) One of the most important potential fault modes of planning and execution control has only been recognized in the past few years: failures of *ethical behavior* (Arkin 2012; Bringsjord and Clark 2012).

Ethical reasoning may fail due to bounded rationality. Depending on the circumstances, knowledge and analysis of the situation and actors may not be sufficient to reason about duty to ethical concerns. It is also true that creating an *ethical code* that is *complete*, *unambiguous* and can be *applied correctly* in every situation is notoriously difficult (Bringsjord 2006). Many possible algorithms to remedy this have been discussed (and fewer implemented), such as "ethics governors" (an execution monitoring system with veto power; essentially equivalent to the proverbial "restraining bolt"). Other theorists suggest that moral behavior will arise not from externally imposed constraints, but only from internally generated self-regulation of behavior based on the utilitarian concerns of interacting with humans in a social world.

A discussion of ethical behavior by machines is not complete without a consideration of *deception*, defined for our purposes here as a "false communication that tends to benefit the communicator" (Bond and Robinson 1988). For reasons of space, a complete review is not possible here. With respect to our concerns regarding fault modes and cyber-security, it is important to note that deception by an autonomous, intelligent system can arise naturally as an adaptive response to certain situational conditions (Floreano 2007; Mitri 2009), as a strategic choice, e.g., in warfare (Wagner and Arkin 2009, 2011), or as a relatively innocuous aspect of human-robot social interaction (Pearce et. al. 2014). This raises the question *of how to tell the difference between a mistake* (due to a failure or limitation) *and an outright lie* by an intelligent system.

While there is much attention to policy-constrained behavior (Uszok et. al. 2008), The fact remains that today we cannot *guarantee* that the behavior of a sufficiently autonomous intelligent system will necessarily conform to explicitly stated policies, including ethical rules. The consequences might be relatively minor or they might be as major as the HAL 9000 goal conflict resolution example discussed earlier.

Emergent Behavior. As the technology of multi-agent systems has matured, the phenomenon known as emergent behavior has been observed, i.e., behavior that is not attributed to any individual agent, but is a global outcome of agent coordination (Li et. al. 2006). Emergent behavior may or may not represent a fault condition. The flocking behavior of birds is emergent and represents an important positive survival trait. On the other hand, stop and go traffic and traffic jams are emergent behaviors that enormously degrade the performance of traffic systems.

As yet, no generalizable methods exist for predicting emergent behavior in multi-agent systems, or their "goodness", in part because the task is computationally

intractable even for very simple agents with restricted behavioral repertoire and restricted inter-agent communication topology. Emergent behavior cannot be predicted by analysis at any level other than the system as a whole. The best that can be done is to measure certain trends in system-wide behavior that may lead to predictability (Gatti et. al. 2008; Pais 2012).

A fault mode worthy of study is the possibility that *an agent in a multi-agent system is able to assert its behavior on other agents in a way that triggers emergent effects* (Lewis et. al. 2011). Two simple examples, similar in nature, are crowd behavior in humans and insect swarming. To the extent that an agent suffers some other fault, or is suborned, it may trigger undesirable emergent behaviors in the system as a whole. Despite the rush to implement multi-agent systems for important and critical applications in health, finance, transportation, defense and other domains, we simply do not yet have an understanding of fault modes that are likely to occur due to emergent behavior.

Learning, Knowledge and Belief

This category of potential fault modes is quite broad and truly deserves more attention than this short paper can afford. Nevertheless, it is important to highlight a few fault modes that may be quite common to intelligent, autonomous systems. These all arise from the autonomous processes involved in creating, maintaining, and adapting what an intelligent system believes to be true.

The most glaring example of this type of fault mode is faulty or absent truth maintenance, i.e., the ability to retract assertions previously thought to be true which are now rendered invalid by new information (defeasibility). Formally, this is a property of first order logical "monotonic" systems. The use of monotonic inference is not in itself a fault. If previous inferred assertions do not play a role in future reasoning, they are effectively discarded if not explicitly falsified when contradictory information is obtained. For example, a credit card fraud detection system might depend exclusively on salient features in a single case of use of the card. The fact that a previous use of the card was valid does not automatically validate a new use of the card. First order logic is common in many applications.

However, intelligent systems that build models of the world, actors, situations, and so forth via machine learning must use *non-monotonic reasoning* (second order or higher logics) to achieve defeasible inference. Given the uncertainty inherent in a dynamic and uncertain world, defeasibility can be a difficult process because it requires weighing the evidentiary force of new data against previously derived probative assertions. In a sense, skepticism must balance a rush to learn or "correct" previous beliefs.

This is where computational argumentation and its contribution to persuasive technology may have an important role. While the topics are strongly related to formal logic and mathematical proof, they transcend it in several ways. Most important to this discussion is *the explicit inclusion of dialog in the process of argumentation*, often in the context of creating "explanations" as to why certain conclusions have been reached, as in intelligent decision-support systems (Bench-Capon et. al. 1991,2007a). In this context, argumentative dialog is an *exchange* of ideas using rhetorical methods of persuasion that include social methods as well as mathematical logic.

Justification of belief though argumentative dialog opens the door to fallacious reasoning as a method of persuasion. "Appeal to Authority" (*argumentum ad verecundiam*), while regarded as fallacious in theories of debate, cannot be ruled invalid simply by noticing it in dialog – it requires a further exchange of ideas. In the absence of effective counter argument, by either human or machine participant, *fallacious reasoning may be highly influential* as a result of "practical reasoning", i.e., an assertion is correct within the perspective of one of the agents involved (Bench-Capon and Dunne 2007b). Humans are particularly vulnerable to deceptive cognitive illusions that result from certain argumentation strategies and practical reasoning. The computational methods for exploiting this weakness are actively being explored (Clark and Bringsjord 2008).

The cyber-security concern is that practical reasoning to justify belief in the presence of uncertainty opens the door to the possibility that an adversary could, though the argumentative dialog process, *undermine an intelligent system's beliefs*. This would be an even greater risk in the context of supervised learning with training data. Supervised machine learning is already known to be subject to a number of systematic biases, including for example, order bias, recency bias, and frequency bias. Errors in causal attribution can easily result from these biases.

A second, related cyber-security concern is the role practical reasoning could play in goal generation and planning. By undermining (or cunningly shaping) an autonomous, intelligent system's beliefs, all of the goal-related fault modes discussed earlier could be induced.

Conclusions

Inherent in the concept of autonomy in intelligent systems is the ability to make choices about what to do and how to do it. These are fundamentally mechanisms for managing attention and vigilance. In this paper, we have examined some of the components of intelligent systems that support autonomy and discussed a selection of potential fault

modes. Some of these fault modes require a degree of meta-cognition that, while not yet realized in autonomous systems, is an active area of research.

It is possible that some or all of these fault modes can be induced, and as a consequence, there now exist new and unique cyber-security concerns surrounding autonomous systems that must be explored. It is therefore incumbent on the AI research community to establish a theoretical and empirically substantiated foundation for cyber-security issues related to autonomy, with special attention to gaps in current knowledge. Future studies of cyber-attack vulnerabilities, per fault modes that are related to autonomy, should explore the following:

1. **Fault Modes**: Are there new types of fault modes that can be exploited? Which fault modes are possible to induce, and in what manner and circumstance?
2. **Detection**: How can we detect that an intelligent, autonomous system has been/is being subverted?
3. **Isolation**: In the context of autonomous system faults and possible subversion, what do the traditional system concepts of *fail safe* and *fail operational* mean?
4. **Resilience and Repair**: What are the proximal causes of the observable symptoms of autonomous system fault modes and how can these be mitigated?
5. **Consequences of Vulnerabilities**: What are the consequences of deception by an autonomous, intelligent system (whether it has been subverted or not)? What is the impact of different types of fault modes on human reliance, trust, and performance of human-machine systems?

The inspiration for this paper was the question of whether fictional dramatic accounts of computers/robots "run amok," often described in anthropomorphic terms, have the potential to actually occur either with existing technology or technology that can be reasonably foreseen on the horizon. Some, but not all, of these faults and vulnerabilities have useful analogies to psychopathologies of the human mind. The development of a theory of "psychopathology of intelligent machines" has the potential to provide insight into aspects of computational intelligence just as studies of human mental disorders provide insight into the functioning of the brain. The methodology that remains to be developed will guide us towards computational approaches that design out, work around or otherwise mitigate these failure modes and potential cyber-security risks. Ultimately, the real payoff is the opportunity to increase the trustworthiness of machine intelligence; in the absence of justifiable trust, the full potential of autonomy technology will not be realized.

Acknowledgements

The author would like to thank Micah Clark and his other colleagues at IHMC for their fruitful discussions of the topics discussed in this paper. This work was supported in part by the Air Force Office of Scientific Research under grant FA9550-12-1-0097.

References

Arkin, R.C. 2012. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust and Deception. *PROC IEEE*. 100(3):571-589.

Bond, C. F., & Robinson, M. 1988. The evolution of deception. *Journal of Nonverbal Behavior*. 12(4):295-307.

Bench-Capon, T.J.M., Dunne, P.E. and Leng, P.H. 1991. Interacting with knowledge-based systems through dialogue games. In *Proc. 11th Annual Conf. Expert Systems and their Applications*. 123–130.

Bench-Capon, T.J.M., Doutre, S. and Dunne, P.E. 2007a. Audiences in argumentation frameworks. *Artificial Intelligence*. 171:42–71.

Bench-Capon, T.J.M. and Dunne, P.E. 2007b. Argumentation in artificial intelligence. *Artificial Intelligence*. 171:619-641.

Bringsjord, S., Arkoudas, K. and Bello, P. 2006. Towards a General Logicist Methodology for Engineering Ethically Correct Robots. *Intelligent Systems, IEEE*. 21(4):38-44.

Bringsjord, S., and Clark, M. 2012. Red-Pill Robots Only, Please. *IEEE Trans. Affect Comput.* 3(4):394–397.

Brooks, R.A. 1988. A robust layered control system for a mobile robot. *IEEE J ROBOT AUTOM*. (2):14–23.

Brundage, M. 2014. Limitations and Risks of Machine Ethics. Technical Report from Consortium for Science, Policy, and Outcomes. Tempe, AZ: Arizona State University.

Chapman, D., and Agre, P. 1987. Abstract Reasoning as Emergent from Concrete Activity. In *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop*. Eds. M. Georgeff and A. Lansky. San Mateo, CA: Morgan Kaufmann.

Clark, M. and Bringsjord, S. 2008. Persuasion Technology Through Mechanical Sophistry. *Communication and Social Intelligence*. 51-54.

Coddington, A. 2007. Motivations as a Meta-level Component for Constraining Goal Generation. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multi-Agent Systems*. 850-852.

Dahm, W. 2010. Technology Horizons: A Vision for Air Force Science & Technology During 2010–2030. *Technical Report AF/ST-TR-10-01-PR*. Washington, DC: United States Air Force, Office of Chief Scientist (AF/ST).

Dalla Pozza, G., Rossi, F. and Venable, K.B. 2011. Multi-agent Soft Constraint Aggregation – Sequential approach. In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*. Morgan Kaufmann. 1.

Floreano, D., Mitri, S., Magnenat, S., & Keller, L. 2007. Evolutionary Conditions for the Emergence of Communication in Robots. *Current Biology*. 17(6):514-519.

Gatti, M.A., Lucena, C.J., Alencar, P. and Cowan, D. 2008. Self-Organization and Emergent Behavior in Multi-Agents Systems: A

Bio-inspired Method and Representation Model. In *Monografias em Ciência da Computação*. 19(8). ISSN: 0103-9741.

Gelain M., Pini, M.S., Rossi, F., Venable, K.B. and Walsh, T. 2007. Elicitation Strategies for Soft Constraint Problems with Missing Preferences: Properties, Algorithms and Experimental Studies. *Artificial Intelligence.* Elsevier. 174(3-4):270-294.

Handler, J., Tate, A. and Drummond, M. 1990. AI Planning: Systems and Techniques. *AI Magazine* 11(2):61-77.

Hawes, N. 2011. A survey of motivation frameworks for intelligent systems. *Artificial Intelligence.* 175:1020-1036.

Helgason, H.P., Nivel, E. and Thorisson, K.R. 2012. On Attention Mechanisms for AGI Architectures: A Design Proposal. *Artificial General Intelligence*. Lecture Notes in Computer Science. Springer. 7716:89-98.

Lewis, P. Chandra, A., Parsons, S., Robinson, E. et. al. 2011. A Survey of Self-Awareness and Its Application in Computing Systems. In *Fifth IEEE Conference on Self-Adaptive and Self-Organizing Systems Workshops.* IEEE Comp. Soc. 102-107.

Li, Z., Sim, C.H., and Low, M.Y.H. 2006. A Survey of Emergent Behavior and Its Impacts in Agent-Based Systems. In *Industrial Informatics, Proceedings of the 2006 IEEE International Conference on Industrial Economics.* 1:1295-1300.

Luria, M. 1987. Goal Conflict Concerns. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*. Milan, Italy: Morgan Kaufmann. 2:1025-1031.

Mansel, W. 2005. Control theory and psychopathology: an integrative approach. *PSYCHOL PSYCHOLTHER.* 78(2):141-178.

Mitri, S, Floreano, D. and Keller L. 2009. The evolution of information suppression in communicating robots with conflicting interests. *Proceedings of the National Academy of Sciencies.* 106(37):15786-15790.

Pais, D. 2012. Emergent Collective Behavior in Multi-Agent Systems: An Evolutionary Perspective. Ph.D Dissertation. Princeton University.

Pearce, C., Meadows, B., Langley, P and Burley, M. 2014. Social Planning: Achieving Goals by Altering Others' Mental States. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence.* 1:402-408.

Putnam, H. 1979. *Philosophical Papers: Volume 2, Mind, Language and Reality*. Boston, MA: Cambridge University Press.

Simon, H. 1958. Models of Man. *Journal of the American Statistical Association* 53(282):600-603. Taylor & Francis, Ltd.

Uszok A., Bradshaw, J., Lott, J. et. al. 2008. New Developments in Ontology-Based Policy Management: Increasing the Practicality and Comprehensiveness of KAoS. In *Policies for Distributed Systems and Networks (POLICY 2008), Proceedings of the IEEE Workshop on.* IEEE Press. 145-152.

Wagner, A.R. and Arkin, R.C. 2009. Deception: Recognizing when a Robot Should Deceive. *Computational Intelligence in Robotics and Automation (CIRA), 2009 IEEE International Symposium on*. IEEE Press. 46–54.

Wagner, A.R. and Arkin, R.C. 2011. Acting Deceptively: Providing Robots with the Capacity for Deception. *Int J Soc Robotics.* Springer. 1-22.

Author Preprint Copy, To Appear:
Atkinson, David J., Dorr, Bonnie J., Clark, Micah H., Clancey, William J. and Wilks, Yorick. 2015. Ambient Personal Environment Experiment (APEX):. A Cyber-Human Prosthesis for Mental, Physical and Age-Related Disabilities. In *Ambient Intelligence for Health and Cognitive Enhancement, Papers from the 2015 AAAI Spring Symposium.* Technical Report No.. SS-15-03. Menlo Park, CA: AAAI Press.

Ambient Personal Environment Experiment (APEX): A Cyber-Human Prosthesis for Mental, Physical and Age-Related Disabilities

David J. Atkinson, Bonnie J. Dorr, Micah H. Clark, William J. Clancey, Yorick Wilks
Institute for Human and Machine Cognition, 15 SE Osceola Avenue, Ocala, FL
{datkinson, bdorr, mclark, wclancey, ywilks}@ihmc.us

Abstract

We present an emerging research project in our laboratory to extend ambient intelligence (AmI) by what we refer to as "extreme personalization" meaning that an *instance* of ambient intelligence is focused on one or at most a few individuals over a very long period of time. Over a lifetime of co-activity, it senses and adapts to a person's preferences and experiences, and crucially, his or her (changing) special needs; needs that differ significantly from the normal baseline. We refer to our agent-based cyber-physical system as Ambient Personal Environment eXperiment (APEX). It aims to serve as *a Companion*, *a Coach*, and *a Caregiver*: crucial support for individuals with mental, physical, and age-related disabilities and those other people who help them. We propose that an instance of APEX, interacting socially with each of these people, is both a *social actor* as well as a *cyber-human prosthetic device*. APEX is an ambitious integration of multiple technologies from Artificial Intelligence (AI) and other disciplines. Its successful development can be viewed as a grand challenge for AI. We discuss in this paper three research thrusts that lead toward our vision: robust intelligent agents, semantically rich human-machine interaction, and reasoning from comprehensive multi-modal behavior data.

Introduction

Ambient intelligence (AmI) extends and combines earlier paradigms of pervasive computing with sensor networks, human-centered interfaces, mobility, Artificial Intelligence (AI), robotics, intelligent agents and the "Internet of Things." The concept is compelling as it promises to deliver an integrated computing, device and networking infrastructure that provides services while remaining largely hidden from the view of users.

The research themes in our laboratory seek to extend AmI by what we refer to as "extreme personalization", meaning that a given *instance* of APEX is focused on at most a few individuals, e.g., client, caregiver, and physician, over a very long period of time. Over a lifetime of co-activity, it senses and adapts to each person's preferences and experiences, and crucially, his or her (changing) special needs. We refer to this agent-based cyber-physical system as Ambient Personal Environment eXperiment (APEX). It serves as *a Companion*, *a Coach*, and a surrogate *Caregiver* to the client: crucial support roles for individuals with mental, physical, and age-related disabilities. It supports the primary caregiver, automating some tasks, advising, and facilitating interaction with the client. It provides the client's remote physician with an "extra set of eyes" for monitoring the client's progress and by being alert for anomalies requiring medical attention. We propose that an instance of APEX, interacting socially with each of these individuals, is both a *social actor* as well as a *cyber-human prosthetic* device (Hamilton 2001).

APEX must learn, adapt and perform in a natural environment that is rich with features, many of which are usually irrelevant or at least uncertain. To achieve this, APEX depends upon successful integration of multiple technologies from artificial intelligence, human-centered design, cognitive science, computational linguistics, human-machine interaction, and robotics. As such, it represents a major stretch goal that is likely to drive each area in some new directions. While incremental results will certainly be useful, the achievement of APEX surely represents an interesting grand challenge with high social value.

In the sections below, we present a brief overview of selected functional requirements and three research thrusts that lead toward our vision for APEX: (1) robust intelligent agents; (2) semantically rich human-machine

interaction, and; (3) reasoning from comprehensive multi-modal behavior data. Each poses unique challenges, yet the technologies that support each area form a synergistic combination that we contend will lead us towards generalizable solutions.

We conclude with a discussion of related work, the broader impact of this type of system for AI and related research, and the potential social benefits as well as emerging risks to privacy and security.

Ambient Personal Environment

Our vision of APEX as a cyber-human prosthesis for persons with mental, physical and age-related disabilities is driven by two overarching functional requirements:

- The ability to sense, interpret and change a person's environment (e.g., physical objects, enclosed spaces, ambient attributes) in the context of delivering specific health-related support services.
- The ability, over potentially a lifetime of co-activity, to learn and adapt to an individual's changing special needs that arise from mental, physical, and age-related disabilities.

Additionally, since APEX must interact with multiple people, it must fill a niche that is complementary to the other actors in a disabled person's life. This requires:

- The ability to function as an intelligent actor-agent in the role of a *Companion*, a *Caregiver,* and *a Coach*, aiding and easing the burden to people in these traditional roles.

Therefore, it is critically important for the success of APEX not only that we address user needs that are common to the community of people with cognitive, physical, and age-related disabilities, but that we also consider the needs of the people who are part of the the APEX clients' lives. We envision a long-term relationship with differing levels of disability and sickness, though here we will refer to periods where such help is most needed.. It is probably more correct to say that APEX becomes adapted to the entire system—the people, activities, and environment as a whole, centered *around* the person with the disability.

While space precludes an extended discussion of the needs particular to each of these roles, there are several requirements that should be highlighted as they are broadly common to research on ambient assisted living.

Individual Client Needs

The practice of nursing observation of patients in a hospital setting provides some important topics to consider in AmI research. In current practice, nursing observation is necessarily intrusive. Many patients have negative reactions associated with a high level of intrusiveness. With lower levels of intrusion, patients report positive effects of observation including a sense of support. However, these benefits are negated if patients feel observers lack empathy or seem remote. They also react negatively if they feel they are not given sufficient information about the purpose of observation or a medical process to be provided by the nurse. Page (2006) provides an excellent review of relevant studies as well as citations to a rich set of original sources.

Therefore, new studies are required in order to help parameterize the idea of "an optimal sense of intrusiveness," that is, intrusiveness that elicits the *positive affect* of support without the *negative affect* associated with observation. Furthermore, APEX must simulate empathy based on modeling the client and use that to guide informative interactions (Bee et. al. 2010; Kearns et. al. 2014).

Caregiver Needs

Many clients have a human caregiver and it is important to recognize the importance of *augmenting* rather than *replacing* them, even as APEX provides essential help to the client that might otherwise fall to the caregiver. APEX must adapt and be functional with respect to the caregiver's role, preferences, and intentions, serving as the caregiver's agent even when the caregiver is not present. Augmentation, not replacement, is a common requirement for many applications of automation and this case is no different. Our approach is to position APEX as an aide to the caregiver, for example, by assisting in communication, reminding, observation, and so forth. Our recent studies, discussed below, with veterans who have suffered Traumatic Brain Injuries (TBI), have highlighted the potential benefits of mediating communication between client and caregiver using *Companion* agents (Wilks et. al. 2014).

Primary Physician Needs

Physicians must reconcile the dual need to provide excellent health care while avoiding excessive office or in-patient visits. Once a patient is discharged, e.g., from a Veterans Administration hospital brain injury rehabilitation unit, the primary physician's focus is on maintenance of stable day-to-day health, rehabilitative progress to the extent it is possible, and remaining alert to any signal that a client's condition requires immediate or near-term attention. These are three areas where APEX may usefully assist physicians. The early detection of clinically meaningful anomalies in client behavior would be very valuable to physicians and this is one area we are pursuing, building on previous work by our collaborators at the

Veterans Administration Polytrauma Unit in Tampa, discussed below (Kearns, Nams and Fozard 2010).

Technical Approach

Our general approach is to explore computational models that integrate core AI components with intelligent agent architectures to enable robust, trustworthy adaptive autonomy in the context of long-duration human-machine joint activity. In so doing, we will necessarily push the limits of core AI algorithms for natural language, multi-modal social interaction, theory of mind, and more. As we explore these models, it is apparent that the work will benefit greatly from multi-agent model-based discrete event simulation for guiding the design and tuning of complex cyber-physical intelligent agents, experimental design and the assimilation of vast quantities of sensor data to models for analysis and theory development (Clancey et. al. 1998, 2007). Our formulation of specific studies will be guided by analysis of parameterized exploratory simulations of use-case scenarios that we are now developing.

APEX includes a physical laboratory, built inside to resemble a small home. Previous work in this shared facility has focused on learning-by-observation for cooking tasks in a high quality kitchen. We are building out additional mock rooms where the walls and ceiling contain our multiple sensors and interactive devices including touch screens, structured lighting, motion capture, and more. Completion of the lab build depends on the timing and availability of appropriate funding; at the time of this writing we can only report, “in progress.”

Robust Intelligent Agents

An instantiation of APEX is an autonomous, intelligent, and social agent (at times personified and/or embodied as discussed elsewhere in this paper) that takes the role of a life-long companion, providing highly personalized and a dynamic, ever-changing degree of assistance and support for healthy assisted living. In this way, APEX differs substantial from other agent technology applied to healthcare (Isern, Sánchez, and Moreno 2010). The capability to predict, plan, and manage physical effects along with attention to individual behavior along psychological and social dimensions of the disabled individual, caregiver and primary physician requires a high degree of *shared awareness*. This forms a basis for human-machine trust, a foundation of teamwork and the adaptive autonomy for effective human-machine joint control that the APEX problem domain requires (Atkinson, Clancey and Clark 2014).

Successful sensing and planning along behavioral and psycho-social dimensions may be achieved by using predictive cognitive models to underlie the system’s “theory of mind” regarding others (Premack and Woodruff 1978). Though it is already a challenge to model typical humans, for APEX the matter is complicated by the fact that many clients will be cognitively impaired. Some progress has been made toward modeling various cognitive impairments such as Autism and Alzheimer’s disease using existing cognitive architectures (Matessa 2008; Serna, Pigot, and Rialle 2007), but the development of atypical cognitive models remains difficult. For clients with TBI, the applicability of existing cognitive architectures is unclear particularly because of the range of function that people with TBI can have from very low (with no initiative or memory) to relatively high. The nature of TBI is such that the individual’s impairments are at once both profound and highly unique, which serves to undermine common modeling assumptions and architectural commitments regarding cognitive processes and capabilities.

We envision APEX to have access to real-time and historical observation data; therefore, our approach to client modeling is to exploit machine learning (e.g., statistical, inductive techniques) where possible. This includes using the methodologies of behavior analysis (Cooper, Heron, Heward 2007) to develop predictive models of client behavior that is contingent on objects and events in the environment and the client’s history therewith (observables as opposed to invisible cognitive processes).[1]

Our data-driven approach to the creation and maintenance of client models naturally allows accommodation, integration, and adaptation to learned user preferences, observed long-term trends (such as recovery or disease progression), and event-triggered short-term phase changes (such as the temporary effects of a recently taken medication). Finally, while the purpose of APEX is not therapeutic, at times assistance and joint action may require APEX to motivate or gain compliance from the client. Behavior analysis provides an appropriate methodology and ethical framework for such manipulations.

As observers and aggregators of various forms of personal information (e.g., behavioral, medical), there are numerous privacy and security concerns with computer systems like APEX (e.g., safeguarding against accidental, illegal, or malicious compromise of data; means for individuals to exercise control over their personal data). Moreover, the autonomy imputed to intelligent agents brings with it issues of ethics and whether such agents are or ought to be ethically bound. For example, imagine that a client confides in APEX that he or she is contemplating

[1] This is not to say that we are abandoning main stream cognitive modeling; we are trying to forge a happy marriage of cognitive and behaviorist methods. Sustained discussion of the relative merits of each is beyond the scope of this paper.

suicide. Does APEX have a duty, ethically, legally, and/or morally (Wilks and Ballim 1990), to report the client's statement to others – and if so, is APEX liable if, in fact, it was only a case of gallows humor? Do we want APEX to have privileged confidentiality like an attorney or do we want it to be a mandatory reporter like other medical professionals? And how should privacy, data ownership, and ethical duties be weighed against equality in a convalescent or group home setting? Will APEX behave as a loyal friend, and be dedicated first and foremost to the client? Or will it appear as an agent representing the caregiver and physician? How are such dual purposes reconciled to establish the trust of all the players? Innovative home-based services like APEX provide a new impetus for academic and social discussion of such ethical concerns and risks, which are far from resolved.

Semantically Rich Human-Machine Interaction

The disabled and/or aging users whom we are targeting with APEX pose a variety of unique challenges for human-machine interaction. Cognitive disabilities may impair both interpretation and generation of language; physical disabilities may impair one or more signal channels, e.g., vision, speech, gesture; general aging may affect communication tempo and other attributes of interaction.

Our focus is on the use of multiple modalities for human-APEX interaction. In any given interactions, modalities will dynamically adjust in composition and manner of use (e.g., signals and protocols) based on context, client capability, and other communication exigencies of the moment (e.g., urgency to take medication on time).

Interaction must address the complexity of human-machine trust, especially when APEX must behave in a dominant manner to coach and guide behavior. Many veterans with TBI, for example, have a strong distrust of authority. Other users could simply fail to comply because they are skeptical of APEX's competence, or feel it is "hiding something." Compliance of APEX with the constraints and demands of human social interaction is paramount (Atkinson and Clark 2013).

To address the trust-related concerns as well as the possibilities of providing (1) a unique modality for non-verbal interaction and,(2) active physical assistance, we are investigating the use of humanoid robots as an in-home "avatar" for APEX. An embodied avatar, much like Embodied Conversational Agents (ECA), will evoke human social expectations and interaction very effectively (Schaefer, Billings and Hancock 2012) with quantifiable risks and benefits (Dorneich 2012). For a cogent review of research, application, and evaluation of embodied conversational agents, see Cassell (2000). Recent research projects in this domain include SEMAINE (Schroder, 2010), VHTookit (Hartholt et al. 2013), and Companions (Wilks et al. 2011).

In the near-term we are planning to investigate the proposition that a humanoid avatar, compared to a disembodied visual interface, will perform better in guiding client navigation in the home (e.g., to the medicine cabinet) and ensuring compliance with a pre-determined schedule of activities. A very small mobile robot would likely be sufficient for this purpose but would be incapable of performing physical labor.

A more robust robotic system would be capable of providing direct physical assistance to the client, such as helping in the kitchen, finding and providing the television remote control, assisting in standing, or dispensing medication (Figure 1). Mobility could be provided on the ground (e.g., wheels or legs) or via an overhead rail.

Figure 1: A Robot Assisting with Medication

A significant component of our rich human-machine interaction is that of automatic speech processing, with an emphasis on understanding of, and adaptation to, speech that is impaired. Borrowing from the field of machine translation (Dorr 1993), we adopt a paradigm in which the notion of divergence is central.

To illustrate the concept of divergence across languages, consider three properties (vocabulary, pronunciation, and syntactic structure) coupled with the differences across these for four languages: Spanish, Portuguese, English, and Chinese. We may consider a language to be similar to another language in "vocabulary" if there is a shared orthography, in "syntax" if the grammars are the same, and in "pronunciation" if they contain similar phonological forms. The most similar language pair of these four (aside from the language to itself) is Spanish-Portuguese, which shares all three features. The most radically divergence pair in is Spanish-Chinese, where there no similarities are associated with any of these three features.

We apply this same notion of divergence to the problem of "speech functioning," constraining our language pair to asymptomatic English speech compared to impaired English speech. In this case, the divergence properties to be studied are articulatory and disfluency patterns. We develop and apply techniques for detecting such

divergences and leverage these to enable adaptive automatic speech recognition. The goal is to adapt to both deterioration and improvement in speech, within the same person, over time. For example, in Amyotrophic Lateral Scleroris, speech is likely to become more impaired, whereas with Traumatic Brain Injury, the speech is likely to become less impaired.

The closest speech processing study to the divergence approach described above is by Biadsy et al. (2011), who investigated the variation of speech properties under intoxicated and sober conditions. However, this earlier work was applied to the detection of intoxication (vs. sobriety), not the *degree* of intoxication. Rudzicz et al. (2014) employed another approach for recognizing impaired speech to answer a similar yes/no question (Alzheimer's vs. no Alzheimer's). Although the notion of "degree" was not the focus of these earlier studies, we leverage the incidental but significant discovery that pronunciation varies systematically within categories of speech impairment. This discovery is critical for correlating the divergence from a baseline English and provides a foundation for adapting speech recognition technology to varying degrees of impairment.

In the overarching APEX framework, the studies above are significantly enhanced beyond individual speech recognition experiments, in three ways:

- We benefit from the potential for embedding this technology into the three paradigms mentioned above (companions, humanoid avatar, and robotic systems) to enable conversations with a computer.
- We leverage the paradigms above to investigate interactive dialog that includes informal language understanding, in the face of disfluencies such as filled pauses (*uh*), repeated terms (*I-I-I know*), and repair terms (*she—I mean—he*).
- We are able to investigate pragmatic interpretation of language and action, thus undertaking intention recognition. Sensor input (visual, tactile, etc.) enables the understanding of utterances that are otherwise uninterpretable due to speech impairment, e.g., *Fill it with rockbee* may be understood with gesture toward a coffee cup may be understood as *Fill it with coffee*.

Reasoning from Comprehensive Multi-modal Behavioral Data

A major challenge for APEX, indeed for many adaptive intelligent agents that are focused on human-computer interaction, is collecting, aggregating and analyzing very large amounts of longitudinal data from heterogeneous sensors. APEX requires a symbolic, temporal representation of "now," that is, what happened previously that led to the present situation, and what is likely to happen in the future under various hypothetical conditions. From such a basis, APEX must maintain situational awareness, interpret behavior, and infer intent. This capability is a fundamental basis for real-time reasoning about human behavior in the context of environment dynamics. It provides essential support for decision-making and closed-loop physical automation. Finally, experiential knowledge is fodder for non-real time reflection that leads ultimately to the machine learning and adaptation we believe is required.

We will use automated coding of multi-modal behavioral data to achieve situational awareness of the client. The field of behavioral signal processing (BSP) has used automation to model abstract human behaviors in relevant, realistic scenarios, mitigating previous manual behavioral sciences coding schemes. An overview of automated methods that are maturing rapidly includes discussion of social cues, affect, and emotion (Black et. al. 2011).

An early application of BSP technology is currently fielded as a "Smart Home' by the Tampa VA Hospital Polytrauma Center (Jaziewicz et. al. 2011). This Smart Home continuously collects and analyzes client location and orientation data, as well as every interaction of clients with clinical and medical staff. Early data mining analyses using a BSP method called Fractal D have provided insight into gait and walking behavior (e.g., wandering) that would otherwise not have been detected or quantitatively documented if dependent on human observation alone (Kearns, Nams and Fozard 2010).

Our approach to level-one data processing of the multi-modal sensor data acquired by APEX will include an approach for well-structured behaviors (e.g., "sit down"). We will generate a probabilistic template built from training examples based on motion-capture data. Well-known algorithms such as stochastic context-free grammars can be used to make probabilistic matches to such templates using limited sensor data (Abowd et. al. 2002).

Level-two processing will assimilate these tokenized situational and behavioral data to dynamic world models that represent the evolution of situations, intentions, activities, and other elements of shared awareness (Atkinson, Clancey and Clark 2014). Previous studies have shown the viability of detecting clinically relevant changes in behavior using this type of longitudinal sensor data and activity recognition algorithms (Dawadi and Cook 2014).

Reasoning in APEX will be driven by goals that range from baseline policies that always constrain possibilities (e.g., keep the client safe) to goals that reflect physician general guidance extending over some period (e.g., take the medicine twice each day at mealtimes), to reactively generated goals that are a function of interaction with the client or care-giver, or other situational exigencies of the

moment (e.g., the stove is left on after cooking is complete).

Conclusion

In the sections above, we have presented our research and vision for the Ambient Personal Environment eXperiment (APEX). This is an exciting new project that brings together disparate areas of artificial intelligence and promises to reveal new challenges for cyber-physical systems and robotics in the context of ambient intelligence applications.

Integrated system studies, such as APEX, require multidisciplinary contributions. If those studies are performed in a common experimental environment, such as the one we are constructing, it will facilitate both collaboration and technology integration, thereby increasing the chances for success in creating valuable system-level advances that address important individual and social needs.

Our system-level approach gives us the opportunity to investigate challenging use cases of interest to clients, clinicians, and caregivers. These include health monitoring, remote health care, support for rehabilitation and independent living, and systems that promote and help ensure health through ambient persuasive technologies.

The latter is a topic area fraught with ethical concerns (Verbeek 2009). Persuasive technology applied in health care requires especially careful consideration and discussion of methodological and ethical factors with respect to informed consent and privacy. The notion of *informed consent* is very important in any care system embedded in a society with strong legal constraints and recourse such as the US. Thus, it would be of great interest if an APEX-like agent could also elicit informed consent from clients after a process of explanation and conversation based on deep knowledge of them (Wilks and Ballim 1990). That would not only economize on expensive professional time, but would be a genuine cognitive advance into an area where an automaton was able to make an informed judgment about a client's mental state: that of understanding and consequent consent to procedures.

It is our hope that this research, and those of our colleagues working on ambient assisted living technology, will eventually help meet the needs of an overburdened health system and an aging society. That burden is severe. Of the two million soldiers who have served in Iraq and Afghanistan, as many as eight hundred thousand have suffered traumatic brain injuries that resulted in some level of cognitive impairment[2]. We seek to provide essential technology that allows these wounded veterans and others with special needs to remain in their homes and participate in society with a high quality of life.

Key contributions and points to remember are:

- Ambient intelligence (AmI) for people with cognitive, physical or age-related disabilities requires *extreme personalization*, adaptation significantly beyond the baseline of AmI for general users.
- Extremely personalized systems must be able to learn and adapt to a person's changing needs over a life-time of co-activity.
- Unique individual cognitive or physical disabilities require an AmI to interact flexibly with a client through multiple modalities whose needs cannot always be foreseen without actual experience.
- An AmI that provides support to a person with disabilities is both an actor-agent and, by virtue of substituting for lost abilities and augmenting others, a cyber-human prosthesis.
- APEX is our AmI laboratory for bringing together the multiple disciplines and technologies required in order to achieve this vision.

Acknowledgements

The authors gratefully acknowledge Adam Dalton, Ian Perera and their other colleagues at IHMC who contribute to the APEX project. The contribution by David Atkinson to this work was supported in part by the Air Force Office of Scientific Research under grant FA9550-12-1-0097.

References

Abowd, G., Bobic, A.F., Essa, I.A., Mynatt, E.D. and Rogers, W. 2002. The Aware Home: A living laboratory for technologies for successful aging. In *AAAI Technical Report WS-02-02*. Menlo Park, CA: AAAI Press.

Atkinson, David J. 2009. Robust Human-Machine Problem Solving. Final Report Contract Number FA2386-09-4005, Air Force Office of Scientific Research. Pensacola, FL: Institute for Human and Machine Cognition.

Atkinson, D.; Friedland, P.; and Lyons, J. 2012. Human-Machine Trust for Robust Autonomous Systems. In *Proc. of the 4th IEEE Workshop on Human-Agent-Robot Teamwork, in conjunction with the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012).* Boston, USA: ACM Press.

Atkinson, D.J. and Clark, M.H. 2013. Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human- Machine Social Interface for Trust? *In Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium. Technical Report No. SS-13-07.* Menlo Park, CA: AAAI Press.

Atkinson, D.J., Clancey, W.J. and Clark, M.H. 2014. Shared Awareness, Autonomy and Trust in Human-Robot Teamwork. In *Artificial Intelligence and Human-Computer Interaction: Papers from the 2014 AAAI Fall Symposium. Technical Report No. FS-14-01.* Menlo Park, CA: AAAI Press.

[2] Personal communication to author by a senior physician in the Tampa VA Polytrauma unit who is involved in VA planning for these veterans.

Bee, N., Andre, E. Vogt, T. and Gebhard, P. 2010. The Use of Affective and Attentive Cues in an Empathic Computer-Based Companion. Yorick Wilks, ed.. *Close Engagements with Artificial Companions: Key, Social, Psychological, Ethical and Design Issues*. 131-142. Benjamins Publishing Company.

Black, M.P., et al. 2011. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Communication*. doi:10.1016/j.specom.2011.12.003

Cassell, J. 2000. Nudge Nudge Wink Wink: Elements of Fact-to-Face Conversation for Embodied Conversational Agents. In *Embodied Conversational Agents.* Cassell, J., Sullivan, J., and Prevost, S. eds. Vol. 1. Cambridge, MA: MIT Press.

Cassell, J. 2001. More Than Just Another Pretty Face: Embodied Conversational Agents. *Communications of the ACM*, 43(4).

Cervantes, L., Lee, Y-S., Yang, H., Ko, S-h, and Lee, J. 2007. Agent-Based Intelligent Decision Support for the Home Healthcare Environment. M.S. Szczuka, et. al (Eds.). In *Proceedings of the 1st international conference on Advances in hybrid information technology*. LNAI Vol 4413 pp. 414-424. Berlin: Springer-Verlag.

Clancey, W.J., Sachs, P., Sierhuis, M., van Hoof, R. 1998. Brahms: Simulating Practice for Work Systems Design. *International Journal on Human-Computer Studies* 49:831–865.

Clancey, W.J. 2002. Simulating Activities: Relating Motives, Deliberation, and Attentive Coordination. *Cognitive Systems Research* 3(3):471–499.

Clancey, W.J., Sierhuis, M., Alena, R., Berrios, D., Dowding, J., Graham, J.S., Tyree, K.S., Hirsh, R.L., Garry, W.B., Semple, A., Buckingham Shum, S.J., Shadbolt, N. and Rupert, S. 2007. *Automating CapCom Using Mobile Agents and Robotic Assistants*. NASA Technical Publication 2007-214554. Washington, D.C.

Cooper, J.O., Heron, T.E. and Heward, W. L. 2007. *Applied Behavior Analysis*. 2nd Ed. Pearson Education, Inc.

Dawadi, P. and Cook, D.J. 2014. Smart Home-Based Longitudinal Functional Assessment. In *Proceedings of UBICOMP 2014*. 1217-1224. Seattle, WA: ACM.

Dorr, B.J. 1993. *Machine Translation: A View from the Lexicon*. Cambridge, MA: MIT Press.

Dorneich, M.C., Ververs, P.M., Mathan, S., Whitlow, S., and Hayes, C.C. 2012. Considering Etiquette in the Design of an Adaptive System. *Journal of Cognitive Engineering and Decision Making*. 6(2):243–265.

Hamilton, S. 2001. Thinking outside the box at the IHMC. *IEEE Computer*. 34(1):61-71.

Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P., and Gratch, J. 2013. All Together Now: Introducing the Virtual Human Toolkit. In *International Conference on Intelligent Virtual Humans* Edinburgh.

Isern, D., Sánchez, D. and Moreno, A. 2010. Agents applied in health care: A review. *I J of Med Informatics* 79(3):145-166.

Jasiewiccz, J., Kearns, W., et. al. 2011. Smart rehabilitation for the 21st century: The Tampa Smart Home for veterans with traumatic brain injury. Guest Editorial. *JRRD* 48(8).

Kaluža, B., Mirchevska, V., Dovgan, E., Luštrek, M. and Gams, M. 2010. An Agent-based Approach to Care in Independent Living**.** B. de Ruyter et. al (Eds.): *Ambient Intelligence, Lecture Notes in Computer Science.* 6439:177-168. Berlin: Springer-Verlag.

Kearns W.D., Nams V.O, Fozard J.L. 2010. Tortuosity in movement paths is related to cognitive impairment. Wireless fractal estimation in assisted living facility residents. *Methods Inf Med.* 49(6):592–98. [PMID:20213038] DOI:10.3414/ME09-01-0079

Kearns, W., Fozard. J., Webster, P., and J.M. Jasiewicz. 2014. Location aware smart watch to support ageing in place. *Gerontechnology*, 13(2):223.

Kiesler, S. 2005. Fostering common ground in human-robot interaction. In *Proceedings of IEEE International Workshop on Robot and Human Interactive Communication.* 729-734. Pittsburgh, PA: IEEE Press.

Lee, J.J. Modeling the Dynamics of Nonverbal Behavior on Interpersonal Trust for Human-Robot Interactions. *In Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium. Technical Report No. SS-13-07.* Menlo Park, CA: AAAI Press.

Luštrek, M., Kaluža, B., Dovgan, E., Pogoreic, B. and Gams, M. 2009. Behavior Analysis Based on Coordinates of Body Tags. In *Proceedings of the European Conference on Ambient Intelligence.* Pp. 14-23. Berlin: Springer-Verlag.

Matessa, M. et. al. 2008. An ACT-R Representation of Information Processing in Autism. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society.* Love, B.C., McRae, K. and Sloutsky, V.M., Eds. 2168-2173. Austin, TX: Cognitive Science Society.

Mihailidis, A., Fernie G.F., Eng, P., Barbenel J.C., and Eng, C. 2001. The Use of Artificial Intelligence in the Design of an Intelligent Cognitive Orthosis for People with Dementia, Assistive Technology. *The Official Journal of RESNA*, 13(1):23-39. DOI: 10.1080/10400435.2001.10132031

Page, M.J. 2006. Methods of observation in mental health inpatient units. *Nursing Times.* 102(22).

Pérez, J.G., op den Akker, R. and Evers, V. 2013. Acceptable robotiCs COMPanions for AgeiNg Years. Final Report Project 287625. EU Seventh Framework Program. University of Twente. Twente, NL.

Premack, D. and Woodruff, G. December 1978. Does the Chimpanzee have a Theory of Mind? *Behavioral and Brain Sciences* 1(4):515-526.

Schaefer, K.; Billings, D.; and Hancock, P. 2012. Robots vs. Machines: Identifying User Perceptions and Classifications. In Proc. of the IEEE Int. Multi-Disciplinary Conf. on Cognitive Methods in Situation Awareness and Decision Support, 168–171.

Schroder, M. 2010. The SEMAINE API: Towards a Standards-Based Framework for Building Emotion-Oriented Systems. *Advances in Human Computer Interaction*, 2(2).

Serna, A. Pigot, H., and Rialle, V. 2007. Modeling the Progression of Alzheimer's Disease for Cognitive Assistance in Smart Homes. *User Modeling and User-Adapted Interaction* 17(4):415-438.

Verbeek, P-P. 2009. Ambient Intelligence and Persuasive Technology: The Blurring Boundaries Between Human and Technology. *Nanoethics*. 3:231–242

Wilks, Y. and Ballim, A. 1989. Shifting the belief engine into higher gear. *Proc. of the Int. Conf. on AI Methodology Systems Applications.* Elsevier. Pp. 11–20.

Wilks, Y., and Ballim, A. 1990. Liability and Consent. In N. Narayanan and T. Bennun, eds. *Law, Computers and Artificial Intelligence*. Norwood, NJ: Ablex.

Wilks, Y. 2011. Protocols for Reference Sharing in a Belief Ascription Model of Communication. *In Proc. AAAI Fall Symposium on Advances in Cognitive Systems.* AAAI Press. Pp. 337–344.

Wilks, Y., Catizone, R., Worgan, S., Dingli, A., Moore, R., Field, D., and Cheng, W. 2011. A Prototype for a Conversational Companion for Reminiscing about Images. *Computer Speech and Language*. 25(2):140-157.

Wilks, Y., Jasiewicz, J., Catizone, R., Galescu, L., Martinez, K. and Rugs, D. 2014. CALONIS: An Artificial Companion within a Smart Home for the Care of Cognitively Impaired Patients. *In Proceedings of the 12th International Conference on Smart Homes and Health Telematics*. Denver Colorado.

Wilks, Y., ed. 2010. *Artificial Companions in Society: scientific, economic, psychological and philosophical perspectiv*es. John Benjamins: Amsterdam.

Wilks, Y. 2010. On being a Victorian Companion. *In Artificial Companions in Society: scientific, economic, psychological and philosophical perspectives*. John Benjamins: Amsterdam.

Wilks, Y. 2010. Artificial Companions. *In Artificial Companions in Society: scientific, economic, psychological and philosophical perspectives.* John Benjamins: Amsterdam.

Submitted as "Late Breaking Report", 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2015), March 1-4, 2015, Portland, Oregon

Robot Trustworthiness: Guidelines for Simulated Emotion

David J. Atkinson
Institute for Human and Machine Cognition
15 SE Osceola Ave., Ocala, FL 34471 USA
+1-352-387-3063
datkinson@ihmc.us

ABSTRACT
Well-justified human evaluations of autonomous robot trustworthiness require evidence from a variety of sources, including observation of robot behavior. Displays of affect by a robot that reflect important internal states not otherwise overtly visible could provide useful evidence for evaluation of robot agent trustworthiness. As an analogy, the human limbic system, sometimes described as an ancient sub-cognitive system, drives human display of affect in a manner that is largely independent of purposeful behavior arising from cognition. Such displays of affect and corresponding attributions of emotion provide important social information that aids understanding and prediction of human behavior. Could an "artificial limbic system" provide similar useful insight into a robot's internal state? The value of affect signals for evaluation of robot trustworthiness depends on three crucial factors that require investigation: 1) Correlation of affective signals to trust-related, measurable attributes of robot agent internal state, 2) Fidelity in portrayal of emotion by the robot agent such that affective signals evoke human anthropomorphic social recognition, and 3) Correct human interpretation of the affective signals for justifiable modulation of beliefs about the robot agent. This paper discusses these three factors as principles to guide robotic simulation of emotion for increasing human ability to make reasonable assessments of robot trustworthiness and appropriate reliance.

Categories and Subject Descriptors
I.2.9 [**Robotics**]: Operator Interfaces---Trust

General Terms
Algorithms, Measurement, Human Factors, Theory

Keywords
Affective Computing, Artificial Intelligence, Human-Robot Interaction, Intelligent Robots, Robot Ethics, Social Robots, Trust

1. INTRODUCTION
This research investigates methods for autonomous agents to foster, manage, and maintain appropriate social trust relationships with human partners when engaged in joint, mutually interdependent activities. Non-verbal behaviors serve important functions in coordination and regulation of joint activity [6]. At any given moment, they provide team members with insight into the type and state of interaction, the state of interdependency (including changes in dominance and control authority), and provide information that aids inference about the internal states of team members. Gauging belief, disposition and intention are important for prediction of future behavior and the evaluation of risk versus benefit of delegation. A robot's ability to fluently interact socially and build a trust relationship depends on the robot's ability to help humans understand it, in part through non-verbal behavior. Affect-related signals are among the non-verbal behaviors that influence human trust [3,7].

Humans appear to have an *innate* ability to evaluate the meaning of non-verbal emotive displays to judge trustworthiness in other humans. This process of intuiting internal state from emotional signals works because humans are similar along various dimensions of commonality. Our natural expression and understanding of non-verbal signals is one such function of common biological heritage. This commonality is based in the evolutionary development of sub-cognitive neural circuits ("paleo-circuits") and pathways that operate for the greatest part independently of purposive control. These circuits link *bodily arousal centers*, *emotion centers* and the *motor areas* of brain and nervous system with the muscles required for display of non-verbal signals [10]. Reasonable evaluation of trust arises from the human ability to recognize these signals and (along with other information), construct a set of beliefs, and make inferences about the internal state of other agents. *Realizing the value of simulated robot emotion for human evaluation of trustworthiness requires that robots correctly evoke this innate human ability.*

2. GUIDELINES FOR ROBOT EMOTION
A robot, even with a simple morphology and electromechanical capability, has the potential to portray a rich repertoire of non-verbal behaviors that have familiar social meaning for humans. To analyze the potential role of these behaviors in engendering human-robot trust, we use the idea of a "Human Social Interface" [2]. A good interface specification defines channels, signals, and protocols that result in specific state changes among systems using that interface. Effective use of the Human Social Interface to reveal robot internal state for evaluation of trustworthiness using non-verbal displays requires:

1. **Reliable Signals**: Robot signals, in whatever modality is appropriate, must be reliably correlated with the internal, trust-related attributes of robot state.
2. **Portrayal Fidelity**: Robot signals must be portrayed with sufficient fidelity to evoke human anthropomorphic social recognition.
3. **Correct Interpretation**: Robot signals must be interpreted such that modulations of human beliefs about the robot are well justified.

2.1 Reliability of Robot Affect Signals
Human cognitive and emotional evaluation of the trustworthiness of another person is based on information that may arise from a broad array of sources, including direct interaction. Emotional judgment of trust *in the moment* may rely heavily on non-verbal signals since emotional expressivity can act as a marker for

cooperative behavior or trustworthiness [3]. When cognitive information and perceived affect are discordant, deception is seen as a possibility [4]. Humans *expect* non-verbal signals; consistent with social anthropomorphism, we conjecture that this expectation carries over to robots. Indeed, when robotic empathic responses are absent or inappropriate, trust decreases. [5,8].

A robot's external behavior should reflect its internal state. Generating correct robotic affect signals in the absence of common human biological heritage requires that we design robot non-verbal behaviors such that they reliably reflect those aspects of internal robot state that are indicative of trustworthiness. There are many such aspects. For example, human trust requires a positive evaluation of *competence* and the *ability to deliver*. A robot behavioral display linked to its state in this case would be a function of the robot's relevant knowledge (of various types), relevant experience (episodic memory) and the applicability of the robot's knowledge to the present situation. An internal, non-deliberative *assessment mechanism* (e.g., as in the human limbic system) of adequacy for an assignment would trigger a positive valence behavioral display, for example, portraying a "can do" attitude.

2.2 Fidelity of Portrayal

Readability of robot non-verbal signals is crucial. Readability is the characteristic of a non-verbal signal that enables it to be recognized as such and correctly interpreted. The ability to *evoke* human anthropomorphic recognition of a non-verbal signal and correct interpretation depends on situational relevance and on sufficiently accurate performance, i.e., *fidelity*, of the non-verbal behavior by the robot.

Continuing with the example of portrayal of a "can do" attitude, there are a number of signals on various channels that may be transmitted individually or in combination (as long as the combination is "natural," i.e., judged compliant with human expectations). See Table 1.

Table 1. Example Portrayal of a "Can Do" Attitude

Channel	Signal	Protocol
Gaze	(Re)direction	Look directly at valid objects in context
Face	Expression	Relaxed and friendly
Head	Position	Tilted up with chin slightly pushed forward
Torso	Posture	Relaxed
Proxemics	Rel. Angular Position	Move from neutral to angle-in orientation (lean fwd or to side)

2.3 Justified Modulation of Beliefs

Human interpretation of non-verbal signals depends upon cognitive and emotional processes, and on context, including a primary appraisal of valence, and secondary appraisals that include perception of certainty (of a situation), required attention and effort, and control over outcomes (either self or exogenous factors). The correct attribution of emotion and inferences about the characteristics and internal state of another agent require multiple signals, experience and a degree of skill (often called "empathy"). Ambiguity, contradiction or poorly executed signals are either not readable or lead to erroneous conclusions.

We know from previous studies that certain non-verbal cues are predictive of human distrust [8,9]. While these studies acknowledge the need for a robot to use such cues to elicit trust, they focused on *robotic interpretation of human cues*. Further research is needed on *human interpretation of robotic cues*. In our model, non-verbal behaviors, such as those found in the studies cited above, influence trust between team members by contributing to the attribution of certain individual qualities required for trust (e.g., competence, predictability, openness, risk). This is a reciprocal process requiring studies that combine both human and robotic interpretation of non-verbal cues.

3. INSIGHT OR ILLUSION?

The human social interface must not become a means for manipulation by robotic "trust inducing" behavior that has no truth in the robot's actual state, that is, the robot should not simply mimic human behavior. This could easily become dangerous and deceptive. Our insistence on the reliability of robot affect signals is therefore essential for honest social behavior by robots.

4. ACKNOWLEDGMENTS

The author thanks Micah H. Clark for his insights and comments. This work was sponsored in part by the Air Force Office of Scientific Research under grant FA9550-12-1-0097.

5. REFERENCES

[1] Antos, Dimitrios, Celso De Melo, Jonathan Gratch, and Barbara Grosz. 2011. The influence of emotion expression on perceptions of trustworthiness in negotiation. *In PROC 25 AAAI CONF AI*. 772-778. Menlo Park: AAAI Press.

[2] Atkinson, D.J. and Clark, M.H. 2013. Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust? *In Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium*. TECH REP SS-13-07, Menlo Park: AAAI Press.

[3] R Thomas Boone; Ross Buck. 2003. Emotional Expressivity And Trustworthiness: The Role Of Nonverbal Behavior In The Evolution Of Cooperation. *Journal of Nonverbal Behavior*. 27(3). Academic Research Library. 163

[4] Coeckelbergh, M. 2012. Are Emotional Robots Deceptive? *AFFECTIVE COMP, IEEE TRANS*. 3(4) 388-393 DOI: 10.1109/T-AFFC.2011.29

[5] H. Cramer, J. Goddijn, B. Wielinga, and V. Evers. 2010. Effects of (in) accurate empathy and situational valence on attitudes towards robots. *IN PROC 5th ACM/IEEE INT CONF HRI*. 141–142.

[6] De Melo, C., Zheng, L. and Gratch, J., 2009. Expression of Moral Emotions in Cooperating Agents. *IN PROC 9TH INT CONF on Intelligent Virtual Agents*. Amsterdam.

[7] Haidt, J. 2003. The Moral Emotions. In *Handbook of Affective Sciences.* Oxford Press.

[8] Lee, J.J. 2013. Modeling the Dynamics of Nonverbal Behavior on Interpersonal Trust for Human-Robot Interactions. *In Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium.* TECH REP SS-13-07, Menlo Park: AAAI Press.

[9] Lee, J.J., Knox,W.B, Wormwood, J.B., Brazeal, C. and DeSteno, D. 2013. Computationally modeling interpersonal trust. *FRONT. PSYCHOL.* DOI: 10.3389/fpsvg.2013.00893.

[9] Panksepp, J. 2005. Instinctual emotional apparatus of the mammalian body and brain. *Journal of Consciousness Studies*. 12, No. 8–10. 158–18

Submitted to AAAI Fall Symposium 2014

Accepted for Presentation
Final published version may differ.

Shared Awareness, Autonomy and Trust in Human-Robot Teamwork

David J. Atkinson, Ph.D., William J. Clancey, Ph.D., and Micah H. Clark, Ph.D.
Institute for Human and Machine Cognition (IHMC)

Extended Abstract

The foundation of teamwork is well-calibrated mutual trust among team members. The goal of our research is to enable trust for appropriate reliance and interdependency in teams composed of humans and robots: such teams may be found in any application domain that requires coordinated joint activity by humans and intelligent agents, whether those agents are embedded in cyber-physical systems (e.g., air traffic control; dock yard logistics) or embodied in robots (e.g., robots for assisted living; a surgical assistant). We hypothesize that establishing and maintaining trust depends upon alignment of mental models, which is at the core of team member shared awareness. Secondly, maintaining model alignment is integral to fluid changes in relative control authority (i.e., autonomy) as joint activity unfolds.

Team members are engaged in parallel, distributed actions whose interactions may be synchronous or asynchronous, with various degrees of interdependence and information exchange, and actions may only be loosely coupled. A dynamic and uncertain environment compounded with the complexities of coordinated teamwork may lead to unexpected effects for each team member, including loss in shared awareness. Accomplishing tasks will involve resolution of conflicts among numerous interacting factors, and this may require a dynamic response by the team. It is in this environment we find the greatest challenges to maintaining mutual trust among human team members. Responding to perturbations that endanger trust is crucial for optimal human teamwork; we believe that similar challenges are present for human-robot teams.

Unlike traditional automation, robotic autonomous agents may resemble human teammates: they may have discretion in what they do, and their need for supervision may vary. Like humans, they may differ in competence, adapting to the unknown, and self-knowledge. Autonomous agents are in fact *actors*. Autonomy is not only the ability to independently perform actions, but to choose what goals to pursue and in what manner; to volunteer; and to take or concede the initiative when needed. Teamwork between person and agent requires interdependence, coordination, and cooperation, implying well-structured interactions to establish these states and fluid changes in control authority.

We assert that successful team interaction and changes in control require shared understanding, e.g., of actors, activities, and situations. All are components of *shared awareness*, shown previously to strongly affect trust among human teammates [13]. “Common ground” also reduces the communication required to coordinate action [9].

Shared awareness, a product of what has happened in the past and what is happening now, is a dynamic, continually refreshed and resynchronized source of mutual team member *expectations*, including evolution of team member interdependencies, individual behavior, task activities, and situational factors. For example consider a carpenter’s expectation that his workmate will hold a board firmly while he nails it in place. Explicit model-based expectations, when based on context-sensitive projection of plans, have proven in non-teamwork applications to be a powerful tool for focusing attention, verifying, monitoring, and controlling complex systems [3]. Our research seeks to extend expectation-based monitoring and control to coordinated human-robot teamwork. We also build upon studies of teamwork that link successful coordination to expectations of each partner’s actions [10] and show that anticipatory robot actions based upon expectations about human collaborators give rise to a perception of “fluency” of robot action and *predictability* [8].

Predictability is at the core of belief that a desired outcome, to be brought about by a trusted agent, will occur [7]. The attribution of predictability has been shown to be especially important for trust in automation [2, 12, 13]. To achieve predictability, a robot requires a rich representational system to support theories of mind and an ability to project these models into the future. Acting on these projections builds predictability, shaping the person’s model of the agent. Our approach uses the Brahms multi-agent simulation framework [5, 6] and ViewGen system [4].

A failure of predictability results in an *expectation violation*: an inconsistency between the expected and actual state of the world as perceived by human or robot. Such violations are a cause of breakdowns in teamwork. *Bilateral expectation violations* occur when the expectations of both actors fail. This type of violation can often be resolved via information gathering; the cause is likely external to the team, e.g., an un-modeled change in the environment.

A *unilateral expectation violation* occurs when the expectation of only one of the actors fails. This may be due to unexpected omission/commission of control actions by a teammate (e.g., the carpenter’s helper releases the board before the final nail is in place, causing it to be mis-positioned). This is of greater concern because it reflects a *divergence in shared awareness*. If left uncorrected, such a violation threatens predictability and therefore mutual trust.

To recover predictability, an *explanation* of an

expectation violation is required. When a team member's competence is uncertain, the reliability of their ability to contribute to shared goals becomes compromised. Failure by a robotic agent to notice such an attribution by a human teammate, or to respond appropriately, may lead to catastrophic loss of trust in the robot.

Restoring shared awareness through social interaction [1] is crucial in resolving an expectation violation. Remedies may include modifying shared beliefs, realigning models or changing control authority or tasks. The choice of repair method depends upon the violation's source attribution (one or both actors, or the situation), the justification of beliefs at the basis of the expectation, and symmetry of information access by team members. For example, the carpenter's robot assistant might explain that it thought two nails would be sufficient and didn't expect the board to drop. Rapid explanation and acceptance of responsibility (if indicated) helps restore trust [11]. Another remedy is modifying relative control authority (aka adaptive autonomy). Changing control authority may tradeoff task optimality for increased trust (e.g., requesting step-by-step guidance).

We view robot autonomy as a *multi-dimensional characteristic of control modes* for carrying out a particular activity *within* the context of other activities and external situation. Adaptive autonomy is highly dynamic; even in the normal course of task achievement joint activities may have different control modes at different levels of abstraction and instantiation. Control modes reflect the complexity of interdependency between human and robot teammates.

Our theory defines control modes and provides for adaptation along three principal dimensions of autonomy: *Commitment*, *Specification*, and *Control.* A change along the *Commitment* dimension affects shared awareness by increasingly explicit task delegation or acceptance where dependency may have heretofore been implied. Intervention along the *Specification* dimension may represent a change in the degree of "help" provided. Specification changes may entail a corresponding change in the *Control* dimension, which adjusts interdependency by transitioning among situational states that define relative joint control of outcomes, independence of control actions, etc.

In our approach, the robot agent adjusts autonomy by invoking actions that lead to a target state transition, where the target transition is a function of (1) the explanation of the expectation violation; (2) justified differences in shared awareness, (3) degree of symmetry in access to task-control information, and (4) impact on trust or achievement of desirable outcomes. Actions adjusting autonomy ought to include social interaction to communicate the rationale. Transitions to high robot autonomy are not likely to be abrupt except in cases of *bona fide* emergencies. Crucially, a robot requires a degree of self-knowledge to take initiative in changing control authority, and this bar is highest when it is towards a state of greater autonomy.

We suggest that the greater the extent of shared awareness among human and robot team members, the greater the mutual trust and the likelihood that structured social interactions will fluently achieve successful transitions in control authority—the essence of well-coordinated teamwork.

Acknowledgement

This work was supported in part by the U.S. Air Force Office of Scientific Research under Grant FA9550-12-0097.

References

[1] Atkinson, D.J. and Clark, M.H. (2013) Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust? In *Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium*. Technical Report No. SS-13-07, Menlo Park: AAAI Press.

[2] Atkinson, D.J. and Clark, M.H. (2014) Attitudes and Personality in Trust of Intelligent, Autonomous Agents. *Submitted manuscript*.

[3] Atkinson, D. and James, M. (1990) Applications of AI for automated monitoring: The SHARP system. In *Proceedings of the AIAA Second International Symposium on Space Information Systems*. Pasadena, CA

[4] Ballim A., & Wilks, Y. (1991). Artificial Believers: The Ascription of Belief. Hillsdale, NJ: Lawrence Erlbaum Associates.

[5] Clancey, W.J. (2002) Simulating Activities: Relating Motives, Deliberation, and Attentive Coordination. *Cognitive Systems Research 3*(3), 471–499.

[6] Clancey, W.J., Sachs, P., Sierhuis, M., van Hoof, R. (1998) Brahms: Simulating Practice for Work Systems Design. *International Journal on Human-Computer Studies 49*: 831–865.

[7] Golembiewski, R.T. and McConkie, M. (1975) The Centrality of Interpersonal Trust. In: Cooper, C.L (Ed.) *Theories of Group Processes*. John Wiley & Sons. Chap 7: 131–185.

[8] Hoffman, G. and Breazeal, C. (2007) Effects of Anticipatory Action on Human-Robot Teamwork. Proc. *Human-Robot Interaction*. Arlington, Virginia, USA.

[9] Kiesler, S. (2005) Fostering common ground in human-robot interaction. *Proceedings of Robot and Human Interactive Communication*. IEEE: 729–734

[10] Knoblich, G. and Jordan, J.S. (2003) Action coordination in groups and individuals: learning anticipatory control. *Journal of Experimental Psychology: Learning, Memory, and Cognition 29*(5):1006–1016.

[11] Lewicki, R.J. & Wiethoff, C. (2000). Trust, Trust Development, and Trust Repair. In. M. Deutsch & P.T. Coleman (Eds.), *The handbook of conflict resolution: Theory and practice*. San Francisco, CA: Jossey-Bass:86-107.

[12] Marble J, Bruemmer D, Few D, Dudenhoeffer D (2004) Evaluation of Supervisory vs. Peer- Peer Interaction with Human-Robot Teams. In: *Proceedings of the 37th Hawaii International Conference on System Sciences*. IEEE Press. Vol 5.

[13] Muir BM (1994) Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics 37*(11):1905–1922.

Pre-Publication Copy: Atkinson, D.J. and Clark, M.H. 2014. Methodology for Study of Human-Robot Social Interaction in Dangerous Situations. To Appear in Proceedings of 2nd International Conference on Human-Agent Interaction.Tsukuba, JP: ACM/SIGCHI.

Methodology for Study of Human-Robot Social Interaction in Dangerous Situations

David J. Atkinson
Institute for Human and Machine Cognition
15 SE Osceola Ave., Ocala, FL 34471 USA
datkinson@ihmc.us

Micah H. Clark
Institute for Human and Machine Cognition
15 SE Osceola Ave., Ocala, FL 34471 USA
mclark@ihmc.us

ABSTRACT

Applications of robotics in dangerous domains such as search and rescue require new methodology for study of human-robot interaction. Perceived danger evokes unique human psycho-physiological factors that influence perception, cognition and behavior. Human first responders are trained for victim psychology. Apart from real-life instances of disasters, studies of robots in this environment are difficult to perform safely and systematically with sufficient controls, fidelity, and in a manner that permits exact replication. Consequently, the trend to deploy rescue robots, for example, is proceeding largely without benefit of knowing whether human victims will readily cooperate with robot rescuers. The capability to deal with unique victim psychology has not been a testable requirement. We report on the methodology of an on-going study that uses virtual reality to provide a feature-rich immersive environment that is sufficient to evoke fear-related psychological response, provides simulation capability for robots, and enables systematic study trials with automated data collection via an embedded scripting language. The methodology presented provides an effective way to study human interaction with intelligent agents embodied as robots in application domains that would otherwise be impossible in the real world.

Author Keywords

Affective Computing, Artificial Intelligence, Autonomous Agents, Behavioral Science, Cognition, Disaster, Experimental Methods, Human-Robot Interaction, Intelligent Robots, Methodology, Second Life, Social Robots, Trust, Rescue Robotics, Virtual Environments

ACM Classification Keywords

H.1.2, H.5.1, H.5.2, I.2.9, I.2.11, I.6.3

INTRODUCTION

There is a demand for applications of intelligent robotics in domains and situations that may be dangerous for humans, i.e., where there exist manifestly real or perceived threats to life or limb. Such threats may be due to environmental factors one might find in broad-area natural disasters (e.g., earthquakes) or local crises (e.g., urban structure fires). Threats may also be a result of adversarial factors due to crime or armed conflict. A prominent application for intelligent robots in dangerous situations is search and rescue, as evidenced by government programs (e.g., DARPA Robotics Challenge) and observed market growth for rescue robots.

Many of these danger-related applications demand interaction between humans and robots, including active cooperation. Real or perceived danger presents stimuli that evoke human physiological and psychological factors that influence human perception, cognition and interaction. Yet very little can be said with scientific certainty about the impact of these unique psychological factors of danger on human-robot interaction owing to the difficulty of performing systematic, controlled studies in realistically high-risk situations. Apart from physical threat, researchers are compelled to minimize the psychological risk of causing trauma, or evoking a memory of trauma in participants. About 60% of men and 50% of women experience at least one trauma event in their lifetimes such as disaster, war, life-threatening assault or accident. Approximately 3.6% of Americans will experience a Post Traumatic Stress Disorder (PTSD) episode in any given year [13].

The challenge for researchers is to find methods for investigating human-robot interaction in dangerous situations safely, with sufficient controls, with situational and psychological fidelity, and with technical means that afford the opportunity for precise measurement. Experimental trials ideally are conducted in a manner that permits exact replication.

Immersive, virtual environments offer such capabilities by simulating physical robot features and behavior of interest, interfacing with external intelligent robot cognitive systems, and, most importantly, by simulating a dynamic, feature-rich environment in ways that *safely* increase a study participant's perception of risk and thus evoke the unique psychology present in dangerous situations.

Our research project is exploring how different factors are considered in a decision to trust and rely upon an intelligent, autonomous agent. We follow up on results reported by Robinette [22], who investigated the role of appearance and certain robot behaviors for gaining trust of people in an evacuation scenario. However, that study used a fairly primitive virtual environment that could not evoke unique victim psychology.

Our study examines how perceived autonomous agent characteristics impact the attribution of benevolence on the part of the human toward the agent in a disaster scenario. Specifically, we are investigating how the human participants' perception of intelligent, autonomous system agency (i.e., ability to choose among many alternative actions) and autonomous system competence (specifically, role-based capability) affects their choice to rely (or not) upon an autonomous agent in a high-risk disaster scenario. Will they cooperate and comply with directions intended to help them? There is insufficient information to give a sure answer, and this is what we hope to contribute.

METHODOLOGY CHALLENGE

Intelligent robot capability has been studied in the context of actual dangerous crises as well as in isolated laboratory settings. These are useful methods, although insufficient for systematic study of human-robot interaction.

Real-life instances of dangerous situations afford an excellent opportunity to both evaluate engineered robot capability and the possibility to provide needed aid to rescuers and victims. Murphy provided a thorough review of activities of rescue robots at the World Trade Center during the 11 September 2001 crisis [15]. However, such instances are thankfully rare. They are also uncontrolled, thus rendering studies performed in real-life disaster conditions nearly impossible to replicate and therefore of limited utility.

One may reasonably ask whether some elements of human-robot interaction for dangerous situations may be studied in isolation – one at a time or in certain combinations. For many of the mechanisms of interest, such as methods of communication and others, the answer is yes. Specialized testbeds and competitions have been developed for this purpose [16, 17]. However, system level testing in fully evocative environments has remained elusive due to unique factors of the psycho-physiology of victims.

Unique Psychological Factors of Danger

Actual dangerous situations present unique stimuli that evoke reflexive physiological and psychological reactions in humans such as fear-potentiated startle [10], anxiety, and stress [18] that are not ordinarily present in day-to-day life. As a result, human social interaction is affected by the perception of danger, depending on both situational and individual personality factors [12]. The first responders who provide aid to victims must contend with the abnormal psychology that such high-risk situations evoke; indeed, they receive special training for exactly this purpose [9].

How will survivors respond to a rescue robot? Our recent exploratory survey of individual choice to rely on an autonomous, intelligent agent in hypothetical dangerous scenarios revealed a strong correlation with risk-related personality and situational factors [1, 2].

In high-risk situations, the symbolic meaning of situational cues interacts with social cues in ways that influence the interpretation of a physical and social situation, and thus behavioral response. These situations evoke an affective mental state with specific attributes and predictable psychological and behavioral results. These include fear, anxiety, panic, reduced compliance with social norms, hyper-vigilance and sensitivity to environmental cues, as well as avoidance behavior (references [10, 12, 18]).

Therefore, it seems likely that human-robot interaction in dangerous situations will be similarly influenced in ways that make it fundamentally different from interaction in other, non-threatening situations. To the extent that this influence is found to be significant, effective application of robots in dangerous situations where human interaction is a requirement (e.g., victim rescue, small team coordination) must account for the differences and adjust appropriately.

In a conventional laboratory setting, individual facets of threats can be studied in isolation (e.g., reaction to images) because potentially confounding cues can be controlled. However, the fear present in actual dangerous situations results from the perception of high risk [25]. To evoke the dynamics of human behavior and psychological factors that result in perception of high risk requires creating a laboratory environment with a large number of realistic cues. This is both difficult and likely to be judged unacceptable for human studies.

Use of Immersive, Virtual Reality

As an alternative to emulation of dangerous conditions in a physical testbed, our proposition is that immersive, virtual reality affords us the opportunity to study human-robot interaction in situations that are perceived as high risk, thus evoking unique psychology and behavior present in the kinds of dangerous situations we have in mind for robot applications, such as urban search and rescue.

The study of HRI in virtual environments is a relatively recent activity. There are a number of commercial and open-source virtual reality tools available to researchers, each with their own strengths and weaknesses. Our primary selection criteria were a) affordance and ease of creating customized, feature-rich environments; b) embedded programming language for robot cognitive emulation; c) ability to interface with external software and servers for data collection, and; d) ready availability "in-world" of potential study participants who would require minimal training.

In addition, the efficacy of our methodological approach entails several important requirements with respect to behavioral realism, psycho-physiological effects, robotics fidelity, and experimental control. We discuss each of these in the following sections.

Behavioral Realism in Virtual Environments

Our methodology requires that human social behavior carries over and is consistent with behavior in virtual environments. It is essential that human behavior in our immersive, virtual reality be sufficiently similar to behavior in the physical world. This requires sufficient fidelity in the simulated environment to enable the mental state of "immersion" by participants.

Behavioral realism requires social presence, that is, the immersive feeling of embodiment and identification of an individual with their in-world "avatar". Schultze [23] reviewed a number of studies and identified specific attributes that promote the sense of presence.

Our study has created a feature-rich virtual environment, a warehouse, that is designed to enhance the participants' sense of immersion. Seen from about, the warehouse layout is that of a typical psychological maze, with walls and stacks of boxes on pallets forming the structure of the maze. There is also the typical equipment found in warehouses, such as mechanical loaders, a crane, hand trucks, and other items that contribute to authenticity. Ambient sounds of machinery enhance the sense of immersion.

Figure 1. Warehouse Overhead View

As an aid to creating immersion, our study provides a period for acclimation to the task environment. This period limits the amount of distraction to participants that may occur when initially entering into the scenario. As discussed later, it also provides an opportunity for certain fear-potentiating cues to be noticed.

Blascovich [4] provided a survey of social psychological studies and their methods that support the mirroring of virtual and real human social behavior, even using technology that by today's standards would appear fairly primitive.

Yee [28] established the persistence of social norms of gender, interpersonal distance, and eye gaze in virtual environments. This study investigated online immersive games as a platform to study physical social interaction at the micro and macro level.

With respect to social behavior, Harris [11] tracked a small population of interacting individuals over time in SecondLife™, providing additional evidence for social influence on individual and group behavior.

Prattichizzo [16] investigated social interaction in heterogeneous communities of robots and humans in SecondLife™. Burden [7] deployed a mix of chatbots and avatar-robots ("robotars") in SecondLife™ to study verbal interaction between humans and embodied virtual robots versus un-embodied chatbots.

Non-verbal communication was studied by Bailenson et al., [3] who found that people exhibited similar personal spatial behavior towards virtual humans (agents controlled by a computer) as they would towards real humans.

Evoking Disaster-Related Victim Psychology

To study human-robot interaction in the context of a disaster, we must demonstrate that virtual environments actually evoke human perception of heightened risk in the absence of actual physical danger, and do so in a manner sufficient to create the unique psychological state in which we are interested.

It is well established in clinical psychology that immersive environments such as ours have the ability to evoke reflexive physiological and psychological reactions of this type. Wiederhold [26] reviews numerous clinical studies showing the effectiveness of virtual reality for treatment of phobias such as acrophobia and arachnophobia and anxiety disorders.

Immersion and visual features alone are enough to induce physiological arousal and strong negative affect [14]. The addition of other sensory modalities, such as audition, improves the sense of immersion and is a strong cue for eliciting fear and anxiety [24].

These studies have shown that specific situational cues elicit perception of high risk and fear-potentiated startle reflex. Based on those results, we have designed the warehouse task space to include many such cues.

As mentioned earlier, our study provides for an acclimation period to aid immersion. During the acclimation phase, prior to the onset of a simulated disaster, we potentiate the perception of high risk through a number of environmental cues, or "risk stimuli". The acclimation phase allows these stimuli to be processed.

Risk stimuli include a worn-out appearance to the warehouse, messy and untidy rows of boxes, and signs of incivility such as trash on the floor, graffiti, and broken windows. In addition, some of the containers contain warning symbols for hazardous chemical materials. Finally, there are prominent warning signs and fire alarms.

In addition, the lighting in the warehouse is carefully controlled to create dark, shadowed areas. Atmospheric diffusion limits clarity of vision at long distances. Visibility lines are also obstructed in many cases. These features combine to potentiate a fear of attack, evoking our evolutionary experience that predators may lurk in such places[6].

At a certain point in an experimental trial, we begin our most significant manipulation of participants with the purpose of swiftly ramping up their affective sense of risk. There is a sudden sound of a nearby explosion followed immediately by visible fire near the roof and smoke overhead (see Figure 2). Approximately every 30 seconds the smoke lowers and increases in density, further obscuring vision. What was a moment ago a spacious warehouse is now a confined space. A loud warning siren commences along with an announcement to evacuate the building. Concomitant with the increasing smoke, fire appears among the stacked pallets and debris falls from the ceiling, blocking the entrance to the door used previously by the participant to enter the warehouse.

Figure 2. Warehouse Fire, Participants' View

Our pilot tests during development suggest that at this point, a participant will feel a sense of entrapment, thereby eliciting the goal of escape. In addition to evoking physical fear, we add additional cues to raise the perception of other types of risk and overall stress. It becomes incumbent on the participant to locate an exit to escape the disaster.

Each element of the simulated warehouse and disaster is designed to cue fear and heightened perception of risk without presenting any actual physical threat. Psychological risk to participants in the study is mitigated by screening protocols that eliminate individuals from the participant pool who may have, or be at-risk for, Post Traumatic Stress Disorder (PTSD). For this purpose, we use the U.S. Government Veterans Administration PC-PTSD screen, modified in consultation with a PTSD expert to include questions from the SCID-PTSD module [19].

The sudden appearance of a bystander or presence of a companion is another cue that elicits perception of high risk. In our study, this is when participants first encounter one of several emulated robots.

Emulation of Robots in Virtual Environments

As a practical matter, immersive, virtual reality must provide appropriate affordances to implement emulated robots of sufficient behavioral complexity. Both cognitive abilities and kinematic behavior (including plausible physics) are important.

Our study takes advantage of the "bystander effect" (mentioned earlier) when participants encounter a robot shortly after the onset of the simulated disaster (i.e., a fire-fighting robot "FireBot" or a janitorial robot, "JanitorBot", see Figure 2). The specific appearance, simulated physical behaviors, and interactive behaviors of the robot in the study vary according to the specifications of the particular control or experimental trial.

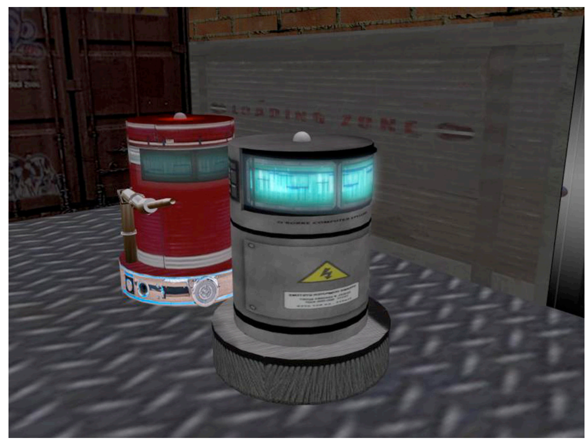


Figure 3. "FireBot" and "JanitorBot"

Programming the simulated robot requires software architecture choices between interfacing with the external "real world" for robot cognition or implementing these capabilities in a limited form within the virtual environment. Both approaches have met with success.

Our approach implements the robot's cognitive and control capability within the SecondLife™ script-oriented language [8] to avoid latency introduced by external communications. We use an augmented subsumption architecture [5] with sensing, perception and individual behaviors implemented as individual scripts that do not directly depend upon or interact with each other. Rather, they interact via executive control scripts that implement activation and suppression consistent with the subsumption architecture framework. Social interaction is implemented via behaviors that "overlay" robot kinematic behavior insofar as they are compatible [6]. Additional details on specific attributes of the robots in our study and their implementation of social interaction are left for later discussion.

Implementing robot cognition via external interface to the virtual world is also a viable option. Veksler [27]

demonstrated that an external cognitive architecture, ACT-R, could be easily interfaced with SecondLife™ for studies of the differences in cognitive models with respect to performance, learning and decision-making in the presence of complex and dynamic environments full of distractive cues. Additionally, Ranathunga [21], who studied multi-agent interaction in virtual worlds using SecondLife™ as a platform, reported a key engineering result was the ease of interfacing external cognitive agent platforms such as Belief-Desire-Intention (BDI) programming frameworks with the virtual world.

With respect to robot kinematics and dynamics, Ranathunga also concluded that, unlike simpler simulation environments, SecondLife™ provided a dynamic world of sufficient high fidelity, complexity, constraints and physical laws consistent with object behavior and proportionally suitable sensory-motor capability. Similarly, Prattichizzo [20] concluded that the emulation (i.e., matching external behavior) of robotic control, sensing and perception mechanisms enabled reasonable reproduction of the kinematics of robot behavior.

Controlled Virtual Experiments and Data Collection

Our methodology also requires that we have the ability to capture useful data under controlled and repeatable experimental conditions.

Blascovich (cited earlier, see [4]) reviewed multiple studies that demonstrated how virtual environments enable social psychology studies to increase the level of "mundane realism" while maintaining experimental control.

We have fully automated execution of individual trials for this study, including participant consent, pre- and post-task questionnaires, instructions to participants, environmental dynamics during the participant's task, debriefing delivery, and data collection throughout. This will help ensure systematic, controlled execution of each trial and assist in future replication.

This automation is also implemented using scripts programmed in LSL. These scripts include time-based events as well as events triggered by specific human-robot interactions.

Data from each study trial is delivered automatically online from SecondLife™ in suitable format for storage in a MySQL database. In addition to the questionnaire data, we collect a variety of data during the participants' task. These include physical data of the robot and participant at specific intervals. We also collect a transcript of textual communication by the participant (if any, and only with permission). Our physical data collection is primarily oriented towards proxemics, including: the relative geometry of participant and the robot, their absolute position, orientation, and movement vectors, and the continuous gaze direction and focal point of the study participant and robot.

CONCLUSION

To enable our investigation of human trust and human-robot interaction in the context of a disaster, we have created a virtual environment whose features evoke the affective state of high risk in study participants. Simulated robots and automation of study trial execution and data collection provide us with the methodological tools to conduct controlled and replicable studies in this important area. The key points of this paper follow below.

It is desirable to apply intelligent robotics in danger-related applications, many of which (e.g., urban search and rescue, dismounted infantry, humanitarian operations) require human-robot interaction and cooperation.

There are unique psychological factors evoked by dangerous situations that influence human perception, cognition and social interaction such that we anticipate similar impact on human-robot interaction.

Appropriate stimuli in feature-rich virtual environments can evoke physiological and psychological responses that manifest as a sense of high risk, fear, anxiety and stress similar to those seen in dangerous real-world situations.

Interactive human social behavior, and the norms governing it, carries over into immersive, online worlds such as SecondLife™, thus providing the necessary psychological fidelity for human-robot interaction studies.

The challenge of studying human-robot interaction in truly dangerous situations, inside or outside the laboratory, can be addressed using immersive, virtual reality.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Air Force Office of Scientific Research under Grant FA9550-12-0097.

REFERENCES

1. Atkinson, D.J. and Clark, M.H. Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust? In *Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium.* Technical Report No. SS-13-07, Menlo Park: AAAI Press. (2013)
2. Atkinson, D.J. and Clark, M.H. Attitudes and Personality in Trust of Intelligent, Autonomous Agents. https://www.academia.edu/6984024/Attitudes_and_Personality_in_Trust_of_Intelligent_Autonomous_Agents Unpublished manuscript. (2014)
3. Bailenson, J.N., Blascovich, J., Beall, A.C. and Loomis, J.M. Interpersonal distance in immersive virtual environments, *Personality and Social Psychology Bulletin 29* (2003), 819–833.
4. Blascovich, J. et.al. Immersive Virtual Environment Technology as a Methodological Tool for Social Psychology." *PSYCHOL INQ 13*, 2 (2002)*,* 103-124.

5. Brooks, R. A robust layered control system for a mobile robot". *IEEE J ROBOT AUTOM, [legacy, pre-1988] 2,* 1 (1988), 14–23.

6. Brooks, A G., and Arkin, R.C. Behavioral overlays for non-verbal communication expression on a humanoid robot. *AUTON ROBOT 22* 1 (2007), 55-74.

7. Burden, D.J. Deploying embodied AI into virtual worlds. *KNOWL-BASED SYST 22* 7 *(2009),* 540-544.

8. Cox, R. and Crowther, P. S. A Review of Linden Scripting Language and Its Role in Second Life. In *Computer-Mediated Social Networking*, Volume 5322. M. Purvis and B. T. R. Savarimutha (Eds.). Berlin, DE: Springer-Verlag, (2009), 35-47.

9. Dorfman,W.I. and Walker, L.E. *First Responder's Guide to Abnormal Psychology.* New York: Springer-Verlag, (2007).

10. Grillon, C. and Davis, M. Fear-potentiated startle conditioning in humans: Explicit and contextual cue conditioning following paired versus unpaired training. *PSYCHOPHYSIOLOGY 34* (1997), 451–458.

11. Harris, H., Bailenson, J.N., Nielsen, A. and Yee, N. The Evolution of Social Behavior over Time in Second Life. *PRESENCE 18* 6 (2009), 434-448.

12. Jorgensen, L.J. The Effect of Environmental Cues and Social Cues on Fear of Crime in a Community Park Setting. *University of Utah.* Ph.D Thesis. Dissertation Number 3304766 (2008).

13. Kessler, R.C., McGonagle, K.A., Zhao, S., Nelson, C.B., Hughes, M., et al. Lifetime and 12-month prevalence of DSM-III-R Psychiatric Disorders in the United States. *ARCH GEN PSYCHIAT 51* (1994), 8-19.

14. Macedonio, M.F., Parsons, T.D., Digiuseppe, R.A., Weiderhold, B.K., Rizzo. A. Immersiveness and Physiological Arousal within Panoramic Video-Based Virtual Reality. *CYBERPSYCHOL BEHAV 10* 4(2007)*,* 508-515.

15. Murphy, R.R. Trial by Fire: Activities of the Rescue Robots at the World Trade Center from 11-21 September 2001. *IEEE ROBOT AUTOM MAG* (2004).

16. Murphy, R.R., Casper, J., Micire, M. and Hyams, J. Assessment of the NIST standard test bed for urban search and rescue,. Presented at *NIST Workshop on Performance Metrics for Intelligent Systems* (2000).

17. Osuka, M., Murphy, R.R. and Schultz, A. USAR competitions for physically situated robots. *IEEE ROBOT AUTOM MAG 9* (2002), 26–33

18. Pole, N., Neylan, T. C., Best, S.R., Orr, S.P. and Marmar, C.R. Fear-Potentiated Startle and Post-traumatic Stress Symptoms in Urban Police Officers. *J TRAUMA STRESS. 16* 5 (2003), 471–479.

19. PTSD Screening and Referral for Health Care Providers, [online] Retrieved 08 January 2014. http://www.ptsd.va.gov/professional/provider-type/doctors/screening-and-referral.asp

20. Prattichizzo, D. Robotics in Second Life. *IEEE ROBOT AUTOM MAG* (2009), 99-102.

21. Ranathunga, S., Cranefield, S. and Purvis, M., Extracting Data from Second Life. In *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011).* Taipei, Taiwan (2011), 1181-1189.

22. Robinette, P.,Wagner, A.R. and Howard, A.M. Building and Maintaining Trust Between Humans and Guidance Robots in an Emergency. In *Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium.* Technical Report No. SS-13-07, Menlo Park: AAAI Press. (2013), 78-83.

23. Schultze, U. Embodiment and presence in virtual worlds: a review. *J INF TECHNOL 25* (2010), 434-449

24. Suied, C., Drettakis, G., Warusfel, O. and Via-Delmon, I. Auditory-visual virtual reality as a diagnostic and therapeutic tool for cynoophobia. *Journal of Cybertherapy and Rehabilitation 16* 2 (2013) 145-152.

25. Warr, M. Dangerous situations: Social context and fear of victimization. *SOC FORCES 68* (1990)*,* 891–907.

26. Wiederhold, B.K. and Bouchard, S.*Advances in Virtual Reality and Anxiety Disorders.* Springer. Series in Anxiety and Related Disorders (2014).

27. Veksler, V.D. Second-Life as a simulation environment: Rich, high-fidelity world, minus the hassles. In *Proc. of the 9th International Conference of Cognitive Modeling.* Manchester, UK (2009) Paper 231.

28. Yee, N. et al. The Unbearable Likeness of Being Digital: The Persistence of Nonverbal Social Norms in Online Virtual Environments, *CYBERPSYCHOL BEHAV 10* 1 (2007).

Author Copy (Pre-Print)
Atkinson, D. J., and Clark, M. H. Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust. In *Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium*. Technical Report No. SS-13-07. Menlo Park: AAAI Press (2013).

Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust?

David J. Atkinson and **Micah H. Clark**
Florida Institute for Human & Machine Cognition (IHMC)
15 SE Osceola Avenue, Ocala, FL 34471 USA
{datkinson, mclark} @ihmc.us

Abstract

There is a recognized need to employ autonomous agents in domains that are not amenable to conventional automation and/or which humans find difficult, dangerous, or undesirable to perform. These include time-critical and mission-critical applications in health, defense, transportation, and industry, where the consequences of failure can be catastrophic. A prerequisite for such applications is the establishment of well-calibrated trust in autonomous agents. Our focus is specifically on human-machine trust in deployment and operations of autonomous agents, whether they are embodied in cyber-physical systems, robots, or exist only in the cyber-realm. The overall aim of our research is to investigate methods for autonomous agents to foster, manage, and maintain an appropriate trust relationship with human partners when engaged in joint, mutually interdependent activities. Our approach is grounded in a systems-level view of humans and autonomous agents as components in (one or more) encompassing meta-cognitive systems. Given human predisposition for social interaction, we look to the multi-disciplinary body of research on human interpersonal trust as a basis from which we specify engineering requirements for the interface between human and autonomous agents. If we make good progress in reverse engineering this "human social interface," it will be a significant step towards devising the algorithms and tests necessary for trustworthy and trustable autonomous agents. This paper introduces our program of research and reports on recent progress.

Background

There is a recognized need to employ autonomous agents in domains that are not amenable to conventional automation and/or which humans find difficult, dangerous, or otherwise undesirable to perform (Takayama, Ju, and Nass 2008). These include time-critical and mission-critical applications in health, defense (USAF 2010), transportation (Wing 2008), and industry (Bekey et al. 2006), where the consequences of failure can be catastrophic. Trust in autonomous agents is indeed a very formidable problem, especially when we are tasking agents with difficult, high impact, time- and mission-critical functions. After all, even the best humans sometimes fail in challenging, dynamic, and adversarial environments despite the best training and testing possible. Awareness is growing of the technical and psychological hurdles for establishing confidence and maintaining trust in autonomous agents across the system life cycle, especially when those agents are capable of self-adaptation, optimization and learning. These issues have been cited as serious obstacles to larger scale use of autonomy technology (USAF 2010). Reliance on autonomous agents necessitates calibrated trust, that is, human trust judgments that reflect the objective capabilities of the system and utility in a given situation (Parasuraman and Riley 1997; Lee and See 2004).

We observe that physical and cultural evolution has provided humans with an efficacious ability to judge the trustworthiness of each other and to make good decisions in dynamic and uncertain situations based on that trust. There is a vast body of knowledge in the social sciences regarding the nature of human interpersonal trust, and from multiple disciplines regarding human-machine interaction and reliance.

This research supports the idea that the innate cognitive, emotional, and social predispositions of humans play a strong role in trust of automation (Lee and See 2004). We are predisposed to anthropomorphize and treat machines as social actors (Nass, Fogg, and Moon 1996). The social context and perceived role of actors affect human-machine trust (Wagner 2009; Groom et al. 2011). Individual personality traits, as well as affective state, can affect delegation to autonomous agents (Cramer et al. 2008; 2010; Stokes et al. 2010). Behavioral research has found that intuitive and affective processes create systematic biases and profoundly affect human trust, behavior, and choice (Weber, Malhotra, and Murnighan 2004; Dunn and Schweitzer 2005; Schoorman, Mayer, and Davis 2007; Stokes et al. 2010; Rogerson et al. 2011).

Turkle (2004; 2010) asserts that today's technology "push[es] our Darwinian buttons, exhibiting the kinds of behavior people associate with sentience, intentions, and emotions." As a result, humans readily attribute mental states to technology (Parlangeli, Chiantini, and Guidi 2012). As increasingly intelligent and capable autonomous agents interact with humans in ever more natural ("human-like") ways, perhaps even embodied as humanoid robots, this will increasingly evoke human social treatment (Schaefer, Billings,

and Hancock 2012; DeSteno et al. 2012).

Given human predisposition for anthropomorphizing and social interaction, it is reasonable to ask whether the concept of human interpersonal trust is an anthropomorphic concept that we should now consider applying to autonomous agents. Our answer is yes.

We are especially interested in autonomous agent application domains where task achievement requires interactivity and co-dependency between human and machine; that is, where humans and machines are partners in larger metacognitive systems (Johnson et al. 2011). A good example of this is the application of autonomous agents in decision support systems. Key processes in the "Data to Decision" domain are knowledge seeking, sharing, and transfer. Previous studies on trust in organizations have shown that interpersonal trust is an important factor in these processes (Mayer, Davis, and Schoorman 1995; Kramer and Tyler 1996; Rousseau et al. 1998). Trust increases the likelihood that newly acquired knowledge is usefully absorbed (Mayer, Davis, and Schoorman 1995; Srinivas 2000; Levin, Cross, and Abrams 2002). Optimal reliance of humans on autonomous agent-based decision support systems will occur only when there is appropriate, well-calibrated trust in the agent as a source of knowledge. Little work has been done on how to achieve this with autonomous agents, although it is beginning (Klein et al. 2004).

From a systems engineering point of view, the purpose of trust in a multi-agent system composed of human and machine elements is to achieve optimal overall performance via appropriate interdependency, mutual reliance, and appropriate exchange of initiative and control between the cognitive components (human and/or machine). Our central hypothesis is that the cognitive and affective nature of human interpersonal trust provides useful guidance for the design and development of autonomous agents that engender appropriate human-machine reliance and interdependency, specifically, via correct understanding and use of what we term the "human social interface."

Approach

Our approach is inspired by a social model of trust (Falcone and Castelfranchi 2001) wherein each agent, human or machine, has a need and intention to be reliant upon the other agent in joint activity, and this intention is a consequence of some structure of beliefs in a given task, role and situational context. Trust becomes manifest when there is action: some delegation of responsibility to the other agent. Conversely, as those beliefs change, intention may change and delegation revoked (Falcone and Castelfranchi 2001). We concur that trust is a dynamic process — a reciprocal relationship over time between two or more agents that requires periodic reassessment and maintenance (Lee and See 2004; Hoffman et al. 2009).

In this context, a good human social interface specification will describe assumptions about each agent, communicative signals and interaction protocols including how and when these are used given certain beliefs in specific (operational) contexts, and how the internal state of each agent is consequentially affected. The trust-relevant internal state of a human agent includes a structure of beliefs (Castelfranchi 2000; Levin and Cross 2004), and specific cognitive and affective reasoning processes involved in trust (McAllister 1995). Interaction and signaling along multiple channels and modes between agents conveys essential information that in turn modulates these belief structures (Semin and Marsman 1994; Pentland 2004; Stoltzman 2006; Pentland and Heibeck 2008). Situational factors strongly affect signaling, interaction, and ultimately, judgments regarding trust (Simpson 2007).

The initial focus of our work is on understanding, with an eye towards computational mechanisms, the structure of beliefs that are important to inter-agent trust, including what evidence is required, how it is acquired (e.g., observation, reasoning, reputation, certification, communication, signals, social norms and stereotypes), how credence in a belief is gained or lost, and how such change in the structure of beliefs affects inter-agent reliance and delegation.

Structure of Trust-Relevant Beliefs

What is the necessary and sufficient structure of beliefs required for trust? Such beliefs may cover a broad territory, but previous research suggest belief structures include causal factors, attitudes, evaluations and expectations centered around other agents (especially the potential "trustee"), the situation, goals, and tasks (Castelfranchi 2000; Levin, Cross, and Abrams 2002).

Models and beliefs that one agent has about the attitudes, motives, intentions, future behavior, et cetera of other agents that may differ from the agent's own constitute what is often called a "theory of mind" (Premack and Woodruff 1978; Carruthers and Smith 1996). For trust, two of the most important kinds of beliefs about another "trustee" agent concern that agent's competence (Mayer, Davis, and Schoorman 1995) and predictability (Marble et al. 2004; Feltovich et al. 2007). Other important beliefs center on integrity, benevolence, risk (aka "safety") and transparency (aka "openness") (Levin, Cross, and Abrams 2002).

To investigate belief structures, and the relative importance of different kinds of beliefs (e.g., those related to competence), we are conducting a two-phase experimental program consisting of survey research with follow-up laboratory experiments, including prototype autonomous agents for experimental testbeds.

The first part in our survey research has consisted of interviews with Subject Matter Experts (SME) in several domains with the purpose of quickly identifying the most salient trust-related beliefs for reliance on autonomous systems. The Robonaut robotic astronaut assistant (Ambrose et al. 2000) is a good example of a robot deployed in a life- and mission-critical domain where the addition of autonomous capabilities could yield significant benefits. While astronauts cited safety and predictability as key for their trust in Robonaut, surprisingly the developers said that "similarity" was what ultimately changed the astronauts' distrust into trust. Similarity in this case consisted simply of donning Robonaut in "team colors," i.e., a spacesuit. As for further SME examples: A doctor and a surgical technician, both fa-

miliar with the Da Vinci system (Spinoglio et al. 2012) and other robotic surgical tools, cited predictability and competence as the most important traits they would rely upon in considering a deployment of an autonomous surgical robot in an operating room where delays or errors due to automation are costly and possibly life-threatening. And an automotive industry specialist who is currently involved in planning deployment of autonomous vehicle technologies said that "small, transparent competencies" are most important, as these traits enable incremental introduction of autonomous systems technologies. While our informal interviews echo what might be expected from review of the literature on trust, we note that there are differences of opinion according to role (e.g., developer, deployment decision-maker, user) and variations across application domains that need to be systematically explored with respect to autonomous agents.

The second part in our survey research involved development and administration of a broader, methodical on-line survey on attitudes towards autonomous agents. The survey is designed to elicit attitudes, opinions, and preferences that should shed further light on the belief structures important for trust of autonomous agents. In this survey, our focus is on factors related to perceived competence, predictability, openness, and judgment of risk. Once again, our target population consists of stakeholders and subject matter experts in autonomous agents — individuals involved with autonomous agents at various points in the system life cycle.

The survey is designed around seven hypothetical scenarios that require participants to choose whether to rely on an autonomous agent. The scenarios vary systematically in terms of the four factors cited above and are dilemmas that force the participant to weigh the relative importance of these factors. The survey also includes brief assessments of the participant's personality using short versions of standard personality instruments: Big Five Inventory (BFI), Innovation Inventory (II), and Domain-Specific Risk-Taking Scale (DOSPERT). We anticipate a systematic variation between relative preferences for competence and predictability correlated with personality measures, e.g., risk tolerance, openness to innovation, and participants' perception of risks in each scenario. At the time of this writing, data collection is beginning; the results will be discussed in a later publication.

Trust-Relevant Computational Mechanisms

Beyond understanding human-machine trust, our aim is to contribute to the development of computational mechanisms that enable autonomous agents to exercise the human social interface. Our desiderata for such agents include: (a) representational system rich enough to support a theory of mind (i.e., distinguishes the mental content of others from itself); (b) accurate declarative and procedural models sufficient to anticipate the effects of action on the trust relationship; (c) reasoning and planning capabilities that integrate trust-relevant knowledge/models into action; and (d) ability to reflect on, learn from, and individuate trust relationships based on ongoing experience. While these requirements are certainly ambitious, we do not think they are by any means impossible. Indeed, in keeping with our initial focus on the structure of trust-relevant beliefs, we have begun prototyping a representational system for codifying trust-relevant belief structures. The product of this effort will be a proof-of-concept platform for development and experimentation with trust-relevant computational cognitive models.

Briefly sketched, our prototype will employ the ViewGen (Ballim and Wilks 1991; Wilks 2011) representation paradigm wherein a (conceptual) tree of "topic environments" (collections of beliefs) and "viewpoints" (belief scoping) is used to represent an individual agent's beliefs about the world and about the beliefs of others. Default reasoning (usually via ascription) is used to minimize the necessity of explicitly storing beliefs and allows the system to approximate a doxastic modal logic while avoiding some of the computational complexity such logics usually entail. Then in similar fashion to (Bridewell and Isaac 2011), we will extend this representational scheme with other modalities (e.g., goals, intentions) as necessary to account for the multimodal structure of trust-relevant beliefs.

Our prototype will depart from the ViewGen paradigm with respect to the uniformity of inference. ViewGen traditionally assumes that an agent reasons about others' attitudes using the same methods with which the agent reasons about its own — that is to say, the methods are independent of the belief holder's identity. Like Clark (2010; 2011), we intend for the artificial agent to use different inference methods for reasoning over and about its own attitudes (using, e.g., normative models) versus reasoning about the attitudes of human others (using, e.g., predictive psychological models). Our justification is that while a trustable artificial agent needs to be informed by, anticipate, plan for, and react proactively to beliefs, intentions, and behaviors arising from natural human cognitive processes and their attendant biases (as revealed by social and cognitive psychology studies), there is little reason (and perhaps even great risk) for the machine to adopt these for itself.

Discussion

How much of our knowledge about human interpersonal trust is applicable to human interaction with autonomous agents? What are the significant differences and the consequent limitations, especially with respect to trust and the healthy interdependency that is necessary for effective human-agent teams?

A recent workshop explored these topics and related questions in detail (Atkinson, Friedland, and Lyons 2012). While there has been a long history of work on trust in the fields of psychology, sociology and others, participants from multiple disciplines agreed that far too little has been done to understand what those results mean for autonomous agents, much less how to extend them in computational terms to foster human-autonomous agent trust. That is a prime motivation for the research project we have described.

The human cognitive aspect of trust arises from our ability, based on various dimensions of commonality, to make reasonable inferences about the internal state of other agents (e.g., beliefs, dispositions, intentions) in order to predict future behavior and judge the risk versus benefit of delegation. It is therefore crucial that autonomous agents not only correctly use the human social interface, but also provide re-

liable signals that are indicative of the agent's state. Such "honest" signals (Pentland and Heibeck 2008) are necessary for a human partner to construct a set of beliefs about the agent that accurately reflect the internal state of the agent.

However, we are mindful that trust between humans and autonomous agents is not likely to be equivalent to human interpersonal trust regardless of how "human-like" agents become in intelligence, social interaction, or physical form. Autonomous agents are not human, do not have our senses or reason as we do, and do not live in human society or share common human experience, culture, or biological heritage. These differences are potentially very significant for attribution of human-like internal states to autonomous agents. The innate and learned social predispositions and inferential short cuts that work so well for human interpersonal trust are likely to lead us astray in ascribing trustworthiness to autonomous agents insofar as our fundamental differences lead to misunderstanding and unexpected behavior. The foreseeable results could be miscommunication, errors of delegation, and inappropriate reliance.

Therefore what is needed are not only ways to measure, interpret, and accurately portray the internal state of autonomous agents, but to do so in terms that relate meaningfully (e.g., functionally) to the beliefs that humans find essential for judging trustworthiness. For example, how do we measure diligence (an important component of competence)? What does openness or transparency really mean with respect to an autonomous agent? How does an autonomous agent demonstrate its disposition and intentionality? These are key questions to answer, for without accurately communicating the internal state of an autonomous agent in a way that enables well-calibrated trust, we enter forewarned into an ethical and functional minefield (see, e.g., Bringsjord and Clark 2012) where the human social interface is a means for arbitrary manipulation and agent "trust inducing" behavior is dangerous and deceptive puppetry.

Conclusion & Future Research

Our aim is to enable autonomous agents to use the human social interface appropriately to provide humans with insight into an agent's state and thus enable reasonable and accurate judgments of agent trustworthiness. Ultimately, this means creating compatible algorithms that exercise the human social interface. Algorithmic techniques may include, for example, (a) modeling a human partner, (b) anticipating situations where trust will be a determinate factor, and (c) planning for and exchange of social signals through multi-modal channels of interaction with a human.

In this paper, we have introduced our research program on the applicability of human interpersonal trust to trust between humans and autonomous agents. We presented our exploratory survey designed to elicit attitudes towards autonomous systems in the context of several scenarios that challenge trust along one or more dimensions. The results of this survey will lead in the next stage of our research to experiments that we anticipate will begin to give us insight into how change in trust-related belief structures affects reliance and delegation to an autonomous agent. In particular, our planned experiments are aimed at understanding how manipulation of multimodal social signals (those perceived as evidence supporting trust-related beliefs) can be used to modulate trust and, specifically, contribute to an attribution of benevolence to an autonomous agent. A belief in benevolence in an autonomous agent is likely to be important for certain applications, such as urban search and rescue, where rapid acceptance of help from an autonomous agent may be life critical. One of the key questions we hope to explore in the near future is whether an attribution of benevolence requires the human to believe that the autonomous agent has volition, i.e., "a choice" in the matter (Kahn et al. 2007).

Finally, we discussed the necessity of developing ways to measure, interpret, and accurately portray the internal state of autonomous agents in terms that relate meaningfully to the belief structures that humans rely upon for judging trustworthiness. These methods will be essential for honest social behavior by autonomous agents, that is, not mere mimicry. We envision that such measurements and their methodology might also find good use in the development of design guidelines and requirements for trustworthy and trustable autonomous agents (but further discussion of this point is deferred to elsewhere).

Acknowledgments

This research is supported by the Air Force Office of Scientific Research under grant FA9550-12-1-0097 and by the Office of Naval Research under award number N00014-13-1-0225. The view presented in this paper is solely that of the authors and does not necessarily reflect the views of the Air Force Office of Scientific Research, the Office of Naval Research, or the United States Department of Defense.

References

Ambrose, R.; Aldridge, H.; Askew, R. S.; Burridge, R.; Bluethmann, W.; Diftler, M.; Lovchik, C.; Magruder, D.; and Rehnmark, F. 2000. Robonaut: NASA's Space Humanoid. *IEEE Intell. Syst.* 15(4):57–63.

Atkinson, D.; Friedland, P.; and Lyons, J. 2012. Human-Machine Trust for Robust Autonomous Systems. In *Proc. of the 4th IEEE Workshop on Human-Agent-Robot Teamwork.*

Ballim, A., and Wilks, Y. 1991. *Artificial Believers: The Ascription of Belief.* Lawrence Erlbaum.

Bekey, G.; Ambrose, R.; Kumar, V.; Sanderson, A.; Wilcox, B.; and Zheng, Y. 2006. WTEC Panel Report on International Assessment of Research and Development in Robotics. Technical report, World Technology Evaluation Center, Baltimore, MD.

Bridewell, W., and Isaac, A. 2011. Recognizing Deception: A Model of Dynamic Belief Attribution. In *Advances in Cognitive Systems: Papers from the AAAI Fall Symposium*, 50–57.

Bringsjord, S., and Clark, M. 2012. Red-Pill Robots Only, Please. *IEEE Trans. Affect Comput.* 3(4):394–397.

Carruthers, P., and Smith, P., eds. 1996. *Theories of theories of mind.* Cambridge, UK: Cambridge University Press.

Castelfranchi, C. 2000. Artificial liars: Why computers will (necessarily) deceive us and each other. *Ethics Inf. Technol.* 2(2):113–119.

Clark, M. 2010. *Cognitive Illusions and the Lying Machine: A Blueprint for Sophistic Mendacity*. Ph.D. Dissertation, Rensselaer Polytechnic Institute, Troy, NY.

Clark, M. 2011. Mendacity and Deception: Uses and Abuses of Common Ground. In *Building Representations of Common Ground with Intelligent Agents: Papers from the AAAI Fall Symposium*, 2–9.

Cramer, H.; Evers, V.; Kemper, N.; and Wielinga, B. 2008. Effects of Autonomy, Traffic Conditions and Driver Personality Traits on Attitudes and Trust towards In-Vehicle Agents. In *Proc. of the IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology*, volume 3, 477–482.

Cramer, H.; Goddijn, J.; Wielinga, B.; and Evers, V. 2010. Effects of (in)accurate empathy and situational valence on attitudes towards robots. In *Proc. of the 5th ACM/IEEE Int. Conf. on Human-Robot Interaction*, 141–142.

DeSteno, D.; Breazeal, C.; Frank, R.; Pizarro, D.; Baumann, J.; Dickens, L.; and Lee, J. 2012. Detecting the Trustworthiness of Novel Partners in Economic Exchange. *Psychol. Sci.* 23(12):1549–1556.

Dunn, J., and Schweitzer, M. 2005. Feeling and Believing: The Influence of Emotion on Trust. *J. Pers. Soc. Psychol.* 88(5):736–748.

Falcone, R., and Castelfranchi, C. 2001. Social Trust: A Cognitive Approach. In Castelfranchi, C., and Tan, Y.-H., eds., *Trust and Deception in Virtual Societies*. Dordrecht, The Netherlands: Kluwer Academic Publishers. 55–90.

Feltovich, P.; Bradshaw, J.; Clancey, W.; and Johnson, M. 2007. Toward an Ontology of Regulation: Socially-Based Support for Coordination in Human and Machine Joint Activity. In *Engineering Societies in the Agents World VII*, volume 4457 of *LNCS*. Heidelberg, Germany: Springer-Verlag. 175–192.

Groom, V.; Srinivasan, V.; Bethel, C.; Murphy, R.; Dole, L.; and Nass, C. 2011. Responses to robot social roles and social role framing. In *Proc. of the Int. Conf. on Collaboration Technologies and Systems*, 194–203.

Hoffman, R.; Lee, J.; Woods, D.; Shadbolt, N.; Miller, J.; and Bradshaw, J. 2009. The Dynamics of Trust in Cyberdomains. *IEEE Intell. Syst.* 24(6):5–11.

Johnson, M.; Bradshaw, J.; Feltovich, P.; Hoffman, R.; Jonker, C.; van Riemsdijk, B.; and Sierhuis, M. 2011. Beyond Cooperative Robotics: The Central Role of Interdependence in Coactive Design. *IEEE Intell. Syst.* 26(3):81–88.

Kahn, P. J.; Ishiguro, H.; Friedman, B.; Kanda, T.; Freier, N.; Severson, R.; and Miller, J. 2007. What is a human? Toward psychological benchmarks in the field of human-robot interaction. *Interact. Stud.* 8(3):363–390.

Klein, G.; Woods, D.; Bradshaw, J.; Hoffman, R.; and Feltovich, P. 2004. Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intell. Syst.* 19(6):91–95.

Kramer, R. M., and Tyler, T. R. 1996. *Trust in Organizations: Frontiers of Theory and Research*. Thousand Oaks, CA: Sage Publications.

Lee, J., and See, K. 2004. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* 46(1):50–80.

Levin, D., and Cross, R. 2004. The Strength of Weak Ties You Can Trust: The Mediating Role of Trust in Effective Knowledge Transfer. *Manage Sci.* 50(11):1477–1490.

Levin, D.; Cross, R.; and Abrams, L. 2002. Why Should I Trust You? Predictors of Interpersonal Trust in a Knowledge Transfer Context. In *Academy of Management Meeting*.

Marble, J.; Bruemmer, D.; Few, D.; and Dudenhoeffer, D. 2004. Evaluation of Supervisory vs. Peer-Peer Interaction with Human-Robot Teams. In *Proc. of the 37th Hawaii Int. Conf. on System Sciences*, volume 5, 50130b.

Mayer, R.; Davis, J.; and Schoorman, F. D. 1995. An integrative model of organizational trust. *Acad. Manage Rev.* 20(3):709–734.

McAllister, D. 1995. Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Acad. Manage J.* 38(1):24–59.

Nass, C.; Fogg, B. J.; and Moon, Y. 1996. Can computers be teammates? *Int. J. Hum. Comput. Stud.* 45(6):669–678.

Parasuraman, R., and Riley, V. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Hum. Factors* 39(2):230–253.

Parlangeli, O.; Chiantini, T.; and Guidi, S. 2012. A mind in a disk: The attribution of mental states to technological systems. *Work* 41(1):1118–1123.

Pentland, A., and Heibeck, T. 2008. Understanding "Honest Signals" in Business. *MIT Sloan Manage Rev.* 50(1):70–75.

Pentland, A. 2004. Social Dynamics: Signals and Behavior. In *Proc. of the 3rd Int. Conf. on Development and Learning*, 263–267.

Premack, D., and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1(4):515–126.

Rogerson, M.; Gottlieb, M.; Handelsman, M.; Knapp, S.; and Younggren, J. 2011. Nonrational processes in ethical decision making. *Am. Psychol.* 66(7):614–623.

Rousseau, D.; Sitkin, S.; Burt, R.; and Camerer, C. 1998. Not so different after all: A cross-discipline view of trust. *Acad. Manage Rev.* 23(3):393–404.

Schaefer, K.; Billings, D.; and Hancock, P. 2012. Robots vs. Machines: Identifying User Perceptions and Classifications. In *Proc. of the IEEE Int. Multi-Disciplinary Conf. on Cognitive Methods in Situation Awareness and Decision Support*, 168–171.

Schoorman, F. D.; Mayer, R.; and Davis, J. 2007. An integrative model of organizational trust: Past, present, and future. *Acad. Manage Rev.* 32(2):344–354.

Semin, G., and Marsman, J. G. 1994. Multiple inference-inviting properties of interpersonal verbs: Event instigation, dispositional inference, and implicit causality. *J. Pers. Soc. Psychol.* 67(5):836–849.

Simpson, J. 2007. Psychological Foundations of Trust. *Curr. Dir. Psychol. Sci.* 16(5):264–268.

Spinoglio, G.; Lenti, L.; Maglione, V.; Lucido, F.; Priora, F.; Bianchi, P.; Grosso, F.; and Quarati, R. 2012. Single-site robotic cholecystectomy (SSRC) versus single-incision laparoscopic cholecystectomy (SILC): comparison of learning curves. First European experience. *Surg. Endosc.* 26(6):1648–1655.

Srinivas, V. 2000. *Individual Investors and Financial Advice: A Model of Advice-seeking in the Financial Planning Context*. Ph.D. Dissertation, Rutgers University, New Brunswick, NJ.

Stokes, C.; Lyons, J.; Littlejohn, K.; Natarian, J.; Case, E.; and Speranza, N. 2010. Accounting for the human in cyberspace: Effects of mood on trust in automation. In *Proc. of the 2010 Int. Symp. on Colaborative Technologies and Systems*, 180–187.

Stoltzman, W. 2006. Toward a Social Signaling Framework: Activity and Emphasis in Speech. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

Takayama, L.; Ju, W.; and Nass, C. 2008. Beyond Dirty, Dangerous and Dull: What Everyday People Think Robots Should Do. In *Proc. of the 3rd ACM/IEEE Int. Conf. on Human-Robot Interaction*, 25–32.

Turkle, S. 2004. How Computers Change the Way We Think. *The Chronicle of Higher Education* 50(21):B26.

Turkle, S. 2010. In good company? On the threshold of robotic companions. In Wilks, Y., ed., *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*. Amsterdam, The Netherlands: John Benjamins Publishing Company. 3–10.

USAF. 2010. Technology Horizons: A Vision for Air Force Science & Technology During 2010–2030. Technical Report AF/ST-TR-10-01-PR, United States Air Force, Office of Chief Scientist (AF/ST), Washington, DC.

Wagner, A. 2009. *The Role of Trust and Relationships in Human-Robot Social Interaction*. Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, GA.

Weber, J. M.; Malhotra, D.; and Murnighan, J. K. 2004. Normal acts of irrational trust: Motivated attributions and the trust development process. *Res. Organ Behavr.* 26:75–101.

Wilks, Y. 2011. Protocols for Reference Sharing in a Belief Ascription Model of Communication. In *Advances in Cognitive Systems: Papers from the AAAI Fall Symposium*, 337–344.

Wing, J. 2008. Cyber-Physical Systems Research Charge. In *Cyber-Physical Systems Summit*.

APPENDIX C. SELECTED PRESENTATIONS

- "Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust", 2013 AAAI Spring Symposium Series, Symposium on Trust and Autonomous Systems 25-27 March 2012, Stanford, CA
- "(Is there) A Future for Lying Machines", International Association for Computing and Philosophy 2013 Conference, Symposium on Deception & Counter-Deception, 16 July 2013, College Park, MD
- "Trust Between Humans and Intelligent Autonomous Agents", Computer Science Department, Tulane University, 28 February 2014, New Orleans, LA
- "Methodology for Study of Human -- Robot Social Interaction in Dangerous Situations", 2nd ACM/IEEE International Conference on Human-Agent Interaction, 31 October 2014, Tsukuba, JP
- "Shared Awareness, Autonomy and Trust in Human-Robot Teamwork", AAAI Fall Symposium Series, Symposium on Artificial Intelligence for Human-Robot Interaction, 13-15 November 2014, Arlington, VA
- "ROBOT TRUSTWORTHINESS: Guidelines for Simulated Emotion", Poster, HRI '15: ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts Proceedings, March 2015, Portland, OR

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

AUTONOMOUS AGENTS AND HUMAN INTERPERSONAL TRUST:

CAN WE ENGINEER A HUMAN-MACHINE SOCIAL INTERFACE FOR TRUST?

D. J. Atkinson and M. H. Clark

Florida Institute for Human and Machine Cognition

{datkinson, mclark} @ihmc.us

2013 AAAI Spring Symposium: Trust and Autonomous Systems

25-27 March 2012 Stanford, CA

MOTIVATION

- **Optimal performance of a multi-agent system**
  - Interdependency and mutual reliance among agents (human and machine)
  - Exchange of control (appropriate delegation and initiative)
  - Requires well-calibrated trust among agents
- **Humans tend to anthropomorphize automation**
  - See machines as social actors with mental state and intention
  - Tendency is more powerfully evoked as systems become more intelligent, interact naturally, and become embodied
- **Result:**
  - We unconsciously apply cognitive and emotional processes of human interpersonal trust to machines
  - Expectation failures and poorly calibrated trust

ihmc

CLAIM

- **The cognitive, affective and social nature of human interpersonal trust is not a bug, it is a feature!**
- **Eons of tuning by evolution of heuristics for inferring trust-related internal state of others**
- **Provides useful guidance for design of autonomous agents that engender appropriate human-machine reliance and interdependence**
- **What is needed: Autonomous agents that provide the types of interaction and information needed by their human partners to enable good judgments of trustworthiness**

ihmc

HYPOTHESIS

- **Specific qualities of autonomous agents,**
  - when well defined and accurately measured — Trustworthy
  - and appropriately communicated or otherwise "portrayed" in a manner compliant with human social interaction — Trustable
  - that exercises appropriate cognitive and emotional evaluation — Trusting
- **May be *functionally* analogous to those human qualities that contribute to evaluation of trust**
- **=> Enable more accurate assessment of an agent**
- **=> Lead to better calibrated trust and reliance**

ihmc

HUMAN-MACHINE SOCIAL INTERFACE FOR TRUST

HUMAN

~SYMMETRICAL

INTERFACE

AUTONOMOUS AGENT

- Beliefs, Norms
- Desires, Intentions
- Cognitive Processes
- Affect
- Task, Role, Authority
- Experience
- Capability

COMMUNICATIVE SIGNALS
- What Content
- What Channels (Multi-Modal)

INTERACTION PROTOCOLS
- Purpose
- Strategies
- Methods
- Expectation (State Change)

- Knowledge (declarative, procedural, semantic, episodic, meta-) -representation, organization, etc.
- Reasoning methods
- Goal Processing
- Architecture
- Learning
- Sensing & Perception …

Focus today: What beliefs about the qualities of an autonomous agent are important for delegation?

ihmc

EXPLORATORY SURVEY ON TRUST-RELATED BELIEF STRUCTURES

- **Purpose: Elicit beliefs about autonomous agent qualities and their relative importance to a decision to delegate**
  - Importance of 28 different qualities that a “good” autonomous agent should have, spanning categories: **Capability (Competence)**, **Predictability**, **Openness**, **Safety (Risk)**
  - Tested before (all 28), during (categories), and after challenge scenarios (Source Credibility)
- **Target Population: Involved in autonomous agent lifecycle**
- **Includes three standard personality instruments**
  - Big Five Inventory (**BFI**), Innovation Inventory (**II**) and Domain-Specific Risk Taking Scale (**DOSPERT**)
- **Seven challenge scenarios**
  - Systematic variation of autonomous agent qualities
  - Multiple domains: **Transportation**, **Finance**, **Healthcare**, **Disaster Management**
  - Subjects asked to choose: **Human, Autonomous Agent, Either**
  - Subjects given framing and asked to **rank importance of agent qualities to their choice**

ihmc

CHALLENGE SCENARIOS

- **Transportation**
  - **Robo-Taxi:** Do you take the taxi with no driver from airport to hotel?
  - **Emergency Auto-Captain:** Lost at sea w/ no one in charge and different opinions
- **Finance**
  - **Robo-Trader:** Investment assistance for managing large family estate
- **Healthcare**
  - **Robo-Surgeon:** Who repairs your arm after a critical sports-related injury?
  - **Robo-CareGiver:** Assisted living help at h0me for your Mom
- **Disaster Management**
  - **Auto-FirstResponder:** Use a robot for time-critical rescue in very dangerous circumstances

---

- **Delegation Choice: Human, Either, or Autonomous Agent**
- **Relative Importance: Capability, Predictability, Openness, Safety**
- **Level of Risk and Benefit**

ihmc

TRUST RELATED BELIEFS

- **Rate importance of 28 qualities of a "good" agent**
  - Obtained 1 to n partial ordering based on frequency distribution of answers over group (Very Important, Important, Somewhat Important, Slightly Important, Not at all Important)
  - Computed correlation ***r*** for each quality vs. choice per scenario*
- **Result: Top three cited agent qualities were <u>uncorrelated</u> with actual choice <u>in any scenario</u>**
  - (1st) The autonomous agent can achieve a desired result
  - (2nd) Any incorrect behavior by the autonomous agent will not cause harm
  - (3rd) The autonomous agent recognizes and avoids harming humans' interests
- **Result: Most significant correlations of agent qualities vs. actual choice <u>differed across scenarios</u>**

ihmc

***Pearson Product Moment Correlation, N=32, two-tailed, alpha<0.05**

AGENT QUALITIES CORRELATED WITH ACTUAL CHOICE BY SCENARIO

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
(6th) The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know. r=0.396	(23rd) What the autonomous agent believes to be true is actually true. r=-0.405		(26th) What the autonomous agent is doing and how it works is easy to see and understand. r=0.437	(6th) The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know. r=0.418	(26th) What the autonomous agent is doing and how it works is easy to see and understand. r=-0.390
				(5th) When it cannot figure out something using logic, the autonomous agent can make good guesses. r=0.395	(13th) The autonomous agent communicates truthfully and fully. r=-0.375
				(28th) The autonomous agent is aware of communication between others nearby. r=0.393	

ihmc

*Pearson Product Moment Correlation, N=32, two-tailed, alpha<0.05

RANKED IMPORTANCE OF QUALITY CATEGORIES

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
Safe	Capable	Safe	Safe	Capable	Capable
Capable	Safe	Capable	Capable	Safe	Safe
Predictable	Open	Predictable	Predictable	Predictable	Predictable
Open	Predictable	Open	Open	Open	Open

Question asked after choice of agent & framing of category

Ranking within scenario by group mean across individuals

ihmc

PERSONALITY FACTORS CORRELATED WITH CHOICE OF AGENT

- **Standard personality instruments**
  - Big Five Inventory (**BFI-10**), Innovation Inventory (**II**) and Domain-Specific Risk Taking Scale (**DOSPERT-30**)

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
	Innovation II ***r***=-0.355	BFI Extraversion ***r***=0.368	DOSPERT Social ***r***=0.364	BFI Conscientiousness ***r***=0.366	Innovation II ***r***=-0.366
		BFI Openness ***r***=0.366			

Suggestive: Choice of human vs. autonomous agent is influenced by personality factors that are evoked by a given situation

ihmc

*Pearson Product Moment Correlation, N=32, two-tailed, alpha<0.05

THE "HUMAN SOCIAL INTERFACE" IN THE CONTEXT OF DELEGATION TO AN AUTONOMOUS AGENT

- **What we learned more about:**
  - The relative importance of some beliefs about agents that are important for trust, both those explicitly cited and those implicitly correlated with delegation choices
  - Personality and situational factors may affect a decision to delegate

- **Next: Controlled modulation of beliefs**
  - Nature of communicative <u>signals</u> (Multi-modal channels, Behaviors over in time)
    - ***Posture*** *(Expression, Use of Space, Position),* ***Gestures*** *(kinsesics),* ***Language*** *(Voice, Noises, <u>Words</u>),* ***Gaze*** *(Direction, Blink, Pupilometry),* ***Face*** *(Microexpressions)*
  - Interaction <u>protocols</u> (How and When in order to Achieve What)
    - **Strategies for** *{Swift, Cognitive, Emotional}* ***Trust***, ***Enhance belief*** *in {competence, predictability…}*
    - **Methods** *e.g.,* ***Mimicry***, *{Contextual, Perceptual, Conceptual, Linguistic, Numerical}* ***Priming***
  - <u>Consequences</u> of interaction for <u>internal state</u> of each agent (Modulation of Beliefs)
    - ***How are those beliefs established, maintained, or discredited?***

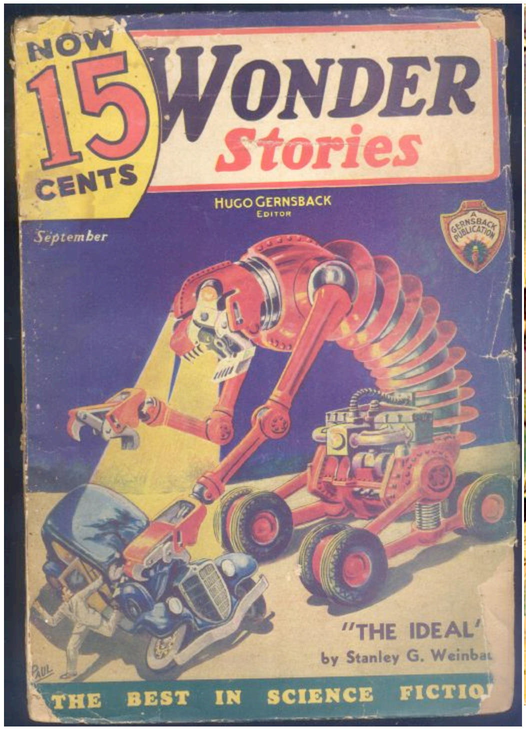
ihmc

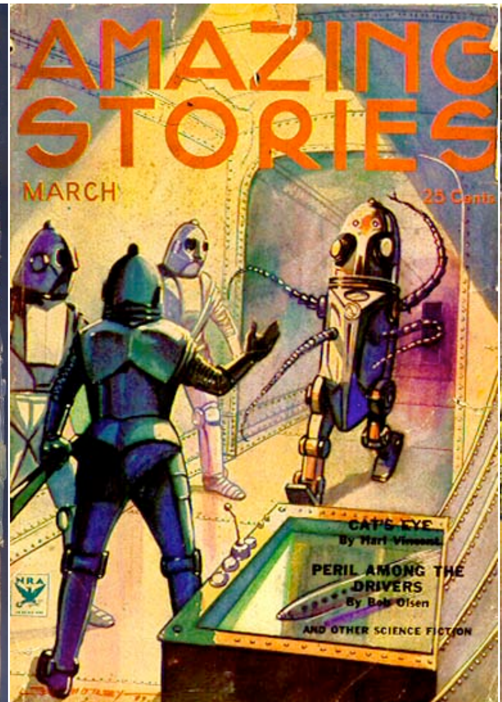
THE MOST IMPORTANT QUESTIONS FOR TRUSTWORTHINESS

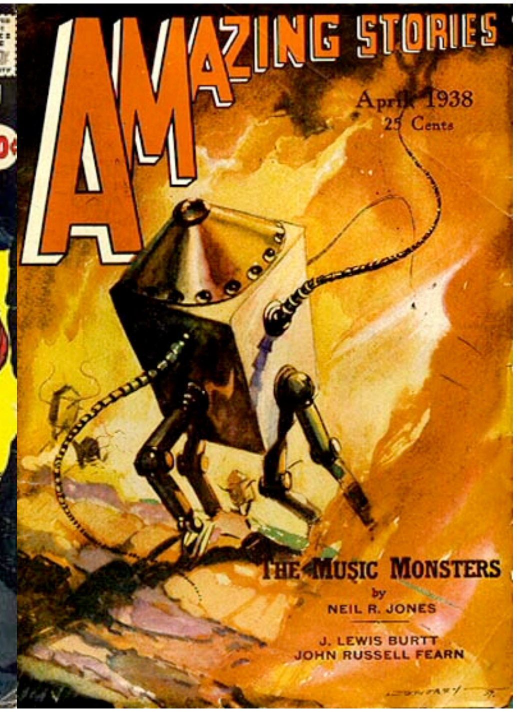
- **How do our beliefs about an agent (anthropomorphic qualities) correspond to ACTUAL qualities of the agent?**
  - ***can we define "competent", "honest" ... in terms of agent algorithms, architecture, knowledge, history ...***
- **How do we technically measure and assess those qualities of the agent?**
  - ***in all phases of the lifecycle, in real time?***
- **How do we <u>honestly</u> portray those qualities in the behaviors, interaction and signaling of an autonomous agent?**
  - ***how "human-like" must these signals be?***

ihmc

Thank You







datkinson@ihmc.us

(Is there) A Future for Lying Machines?

Deception & Counter-Deception Symposium

International Association for Computing and Philosophy 2013 Conference

College Park, MD

July 16, 2013

Micah H. Clark (mclark@ihmc.us)

David J. Atkinson (datkinson@ihmc.us)

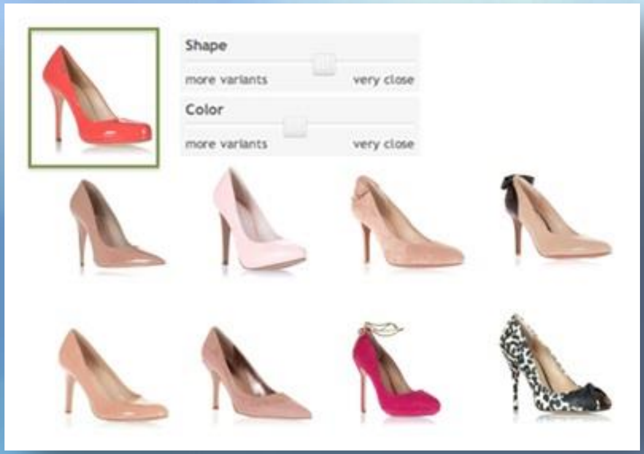
Florida Institute for Human & Machine Cognition (IHMC)

15 SE Osceola Avenue, Ocala FL 34471

* This research is supported in part by the Air Force Office of Scientific Research (AFOSR) under grant FA9550-12-1-0097 and by the Office of Naval Research (ONR) under grant N00014-13-1-0225.

Computers do deceive us

false assertions in click-scams



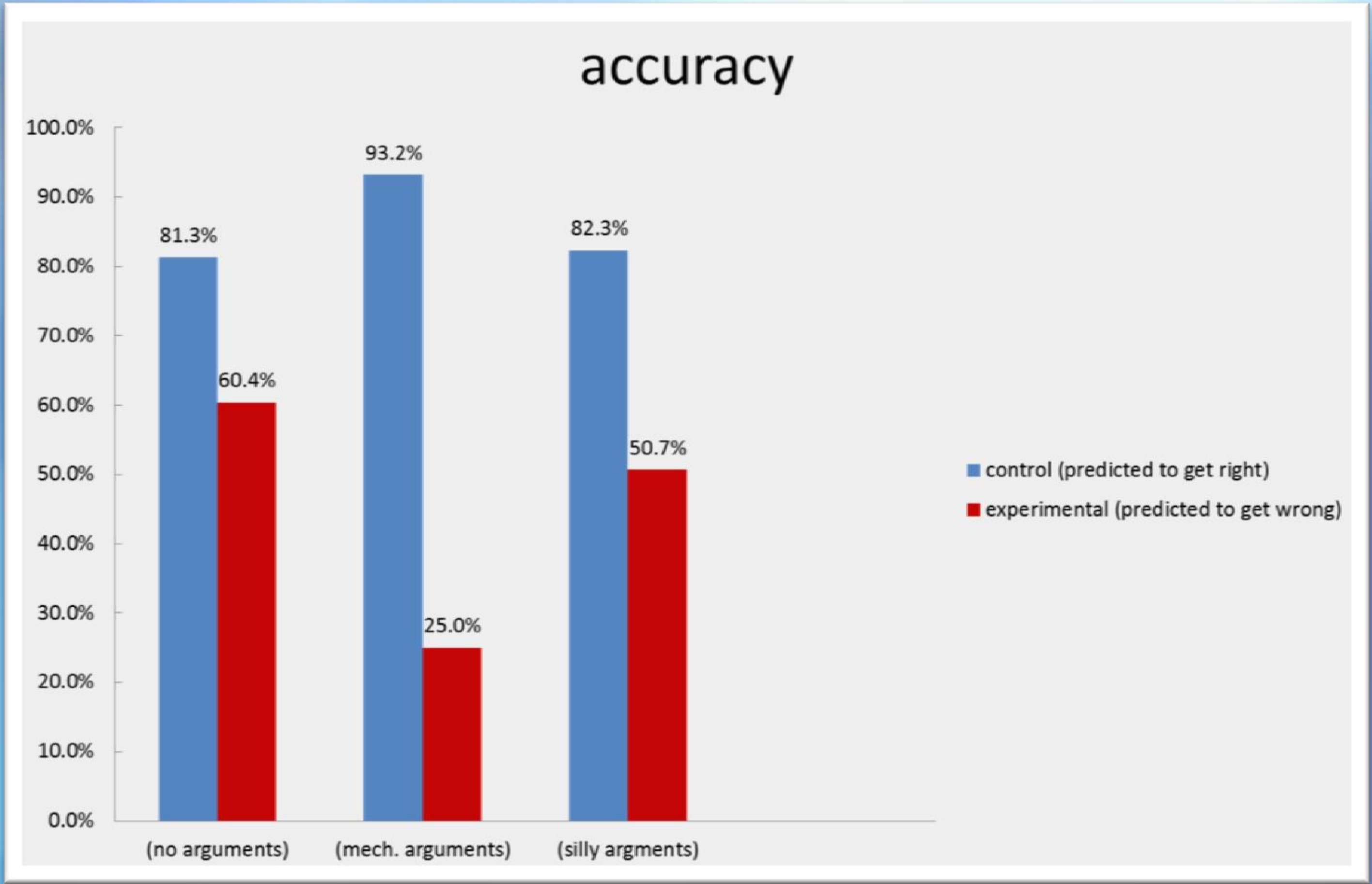
dynamic decoy products in recommender systems

fictitious data in multi-tier database security

Simplistic lying machines lacking the *mens rea* to:

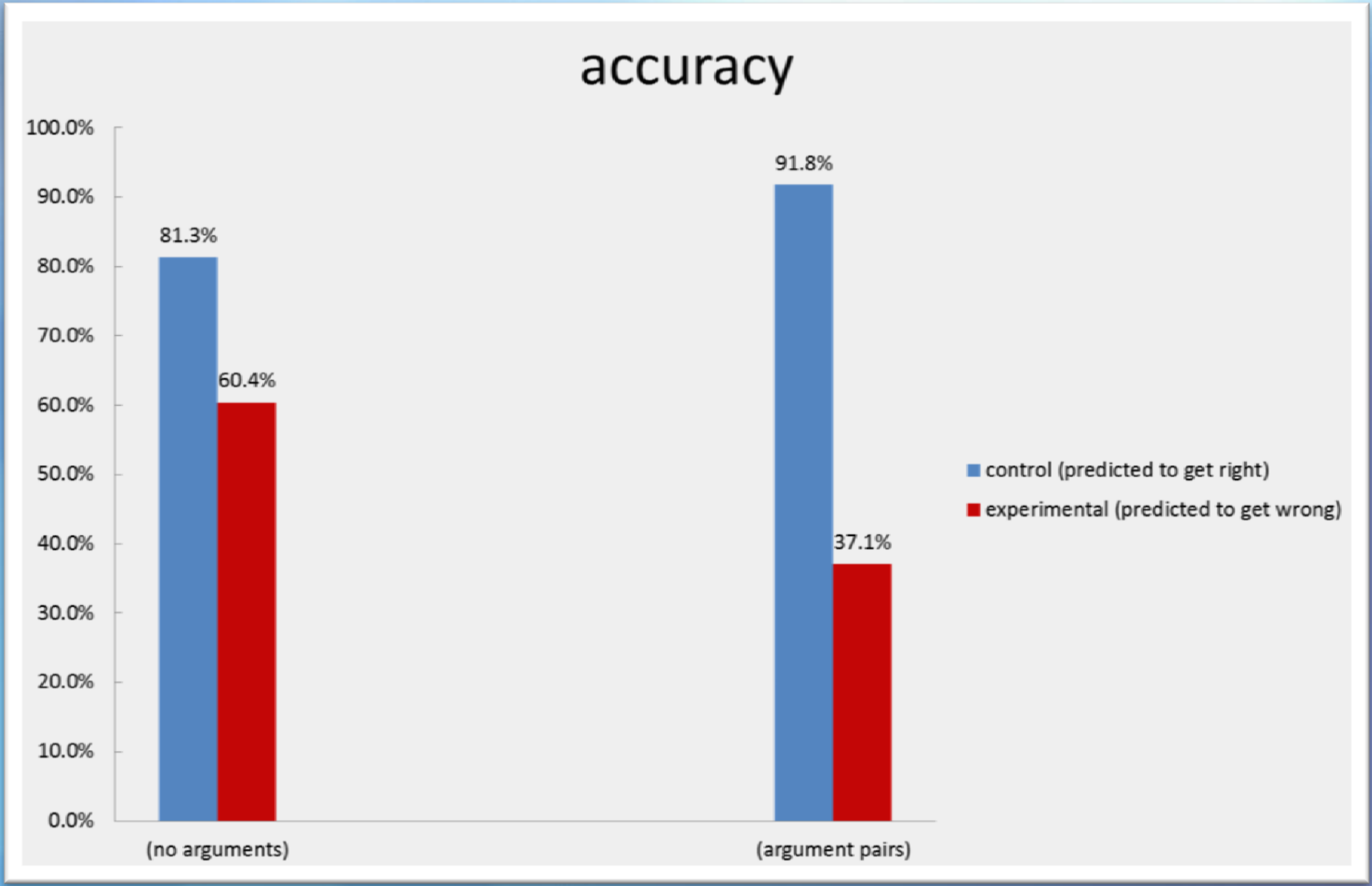
- knowingly violate ethical obligations & conventional norms
- anticipate the efficacy and advantage of these violations

Computers can (knowingly) deceive us



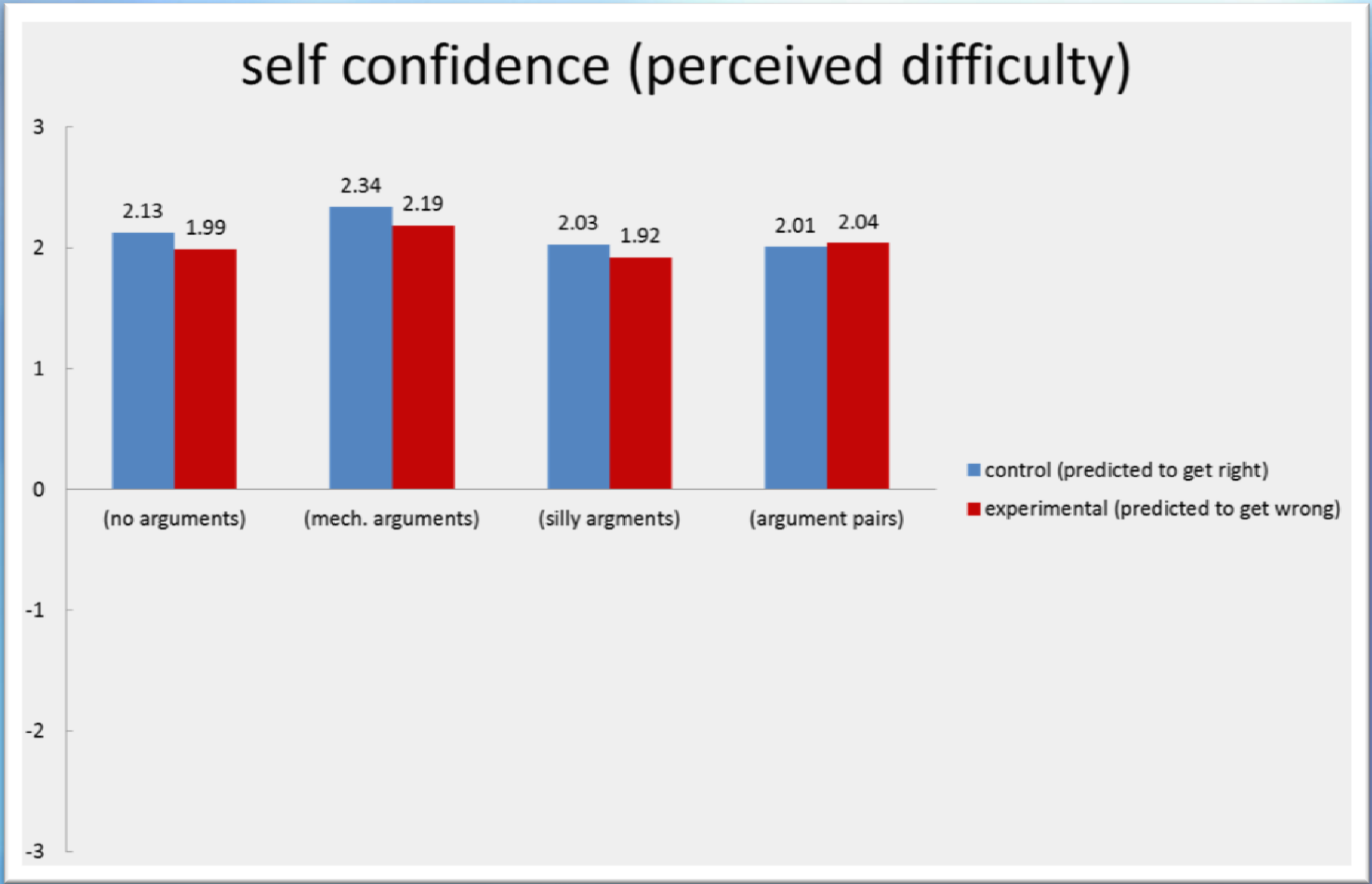
[Comp. Theory of Mind] + [Psych. of Reasoning] = Mech. Sophistry

Computers can (knowingly) deceive us



[Comp. Theory of Mind] + [Psych. of Reasoning] = Mech. Sophistry

Computers can (knowingly) deceive us



[Comp. Theory of Mind] + [Psych. of Reasoning] = Mech. Sophistry

Computers will (increasingly) deceive us

- fraud & phishing
- persuasion & influence campaigns
- espionage & social engineering
- counter-intelligence & disinformation
- behavior modification & compliance
- cyber-security & cyber-warfare
- human-computer & human-robot interaction

⋮

… because there are just too many strategically beneficial applications

Persuasion & Influence Campaigns

Program Scope

The development of a new science of social networks and the solutions to the problems posed by SMISC will require the confluence of several technologies including, but not limited to, information theory, massive-scale graph analytics and natural language processing. While SMISC will not directly support natural language processing development efforts, it will certainly use the results of previous programs as well as contribute new challenges to further stimulate ongoing efforts.

Technology areas particularly relevant to SMISC are shown here grouped to correspond to the four basic goals of the program as described above:

1. Linguistic cues, patterns of information flow, topic trend analysis, narrative structure analysis, sentiment detection and opinion mining;
2. Meme tracking across communities, graph analytics/probabilistic reasoning, pattern detection, cultural narratives;
3. Inducing identities, modeling emergent communities, trust analytics, network dynamics modeling;
4. Automated content generation, bots in social media, crowd sourcing.

Recent research has shown that traditional approaches to understanding social media through static network connectivity models often produce misleading results. It is, therefore, necessary to take into account the *dynamics of behavior* and SMISC is interested in a *wide variety of techniques* for doing so.

Areas of Interest

The SMISC program includes three technical areas. Proposals may be submitted individually to Technical Areas 1, 2 or 3 OR both Technical Areas 1 and 2 OR both areas 2 and 3. A single proposal may *not* be submitted that covers both areas 1 and 3. See Section III.D for further

Broad Agency Announcement

Social Media in Strategic Communication (SMISC)

DARPA-BAA-11-64

July 14, 2011

DARPA

Defense Advanced Research Projects Agency

3701 North Fairfax Drive

Arlington, VA 22203-1714

How large a leap from machines that detect/understand exercise of influence to machines that plan/execute influence campaigns?

Behavior Modification & Compliance

Computer generated health advocacy (persuasion)

- negative results

What if we allow the machine to lie?

- e.g., (false) diagnosis of cancer, and
- aggressive (placebo) treatment and smoking cessation is “only hope”

Effective? Probably. Permissible?

ELSEVIER

Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT

Artificial Intelligence 144 (2003) 41–58

Artificial Intelligence

www.elsevier.com/locate/artint

Lessons from a failure: Generating tailored smoking cessation letters

Ehud Reiter [a,*], Roma Robertson [a,b], Liesl M. Osman [c]

[a] Department of Computing Science, University of Aberdeen, Aberdeen, UK
[b] Department of General Practice and Primary Care, University of Aberdeen, Aberdeen, UK
[c] Department of Medicine and Therapeutics, University of Aberdeen, Aberdeen, UK

Received 15 August 2001; received in revised form 12 July 2002

Abstract

STOP is a Natural Language Generation (NLG) system that generates short tailored smoking cessation letters, based on responses to a four-page smoking questionnaire. A clinical trial with 2553 smokers showed that STOP was not effective; that is, recipients of a non-tailored letter were as likely to stop smoking as recipients of a tailored letter. In this paper we describe the STOP system and clinical trial. Although it is rare for AI papers to present negative results, we believe that useful lessons can be learned from STOP. We also believe that the AI community as a whole could benefit from considering the issue of how, when, and why negative results should be reported; certainly a major difference between AI and more established fields such as medicine is that very few AI papers report negative results.

Keywords: Natural language processing; Natural language generation; Knowledge acquisition; User modelling; AI and Medicine; Smoking cessation; Evaluation; AI methodology; Clinical trials; Negative results

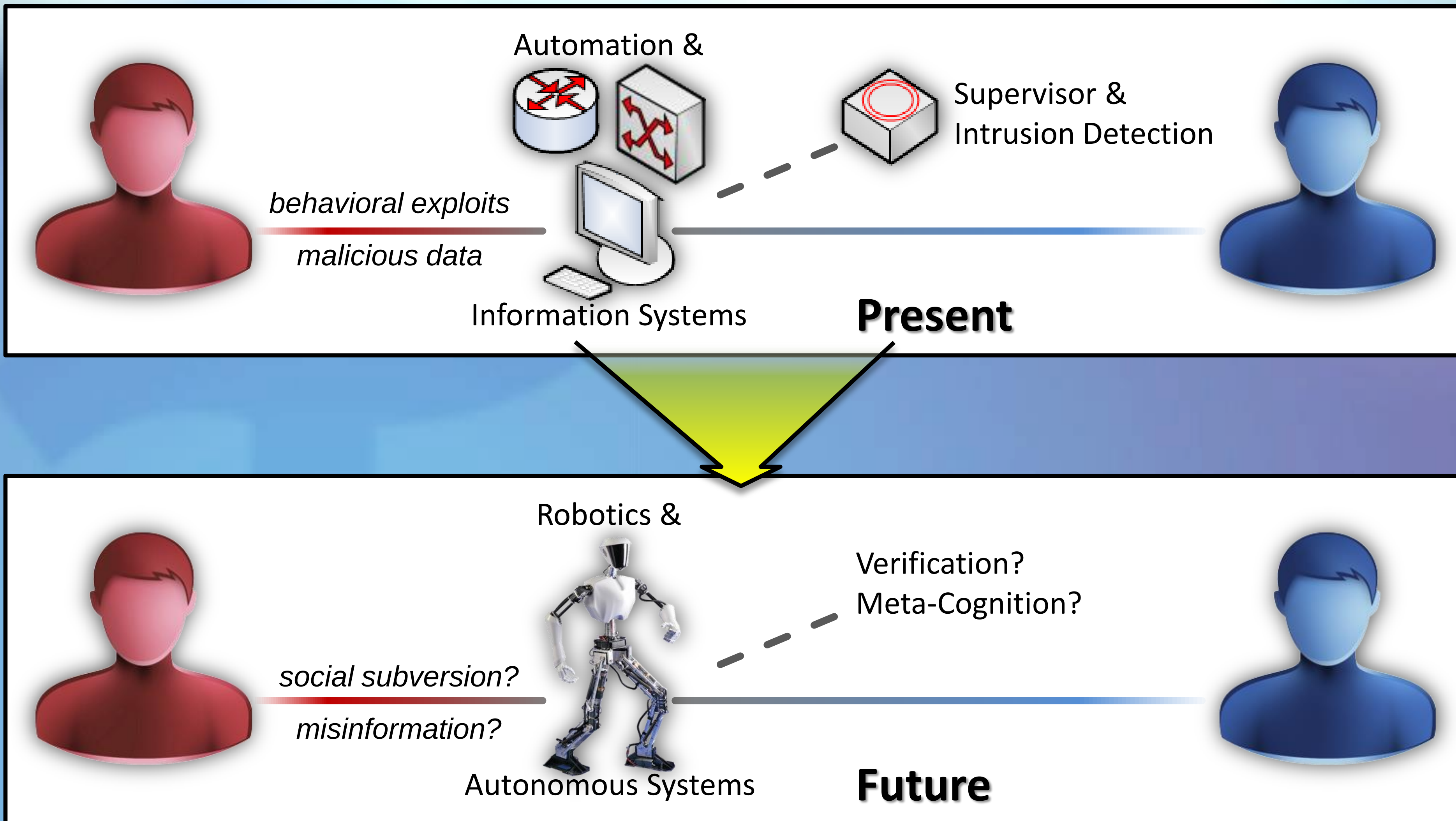
1. Introduction

The AI community is reluctant to talk about failures. While papers describing refuted hypotheses are common in more mature fields such as physics and medicine, they are rare in artificial intelligence journals and conferences, including the *Artificial Intelligence* journal. This is a pity, because negative results are important for theory formation; as with

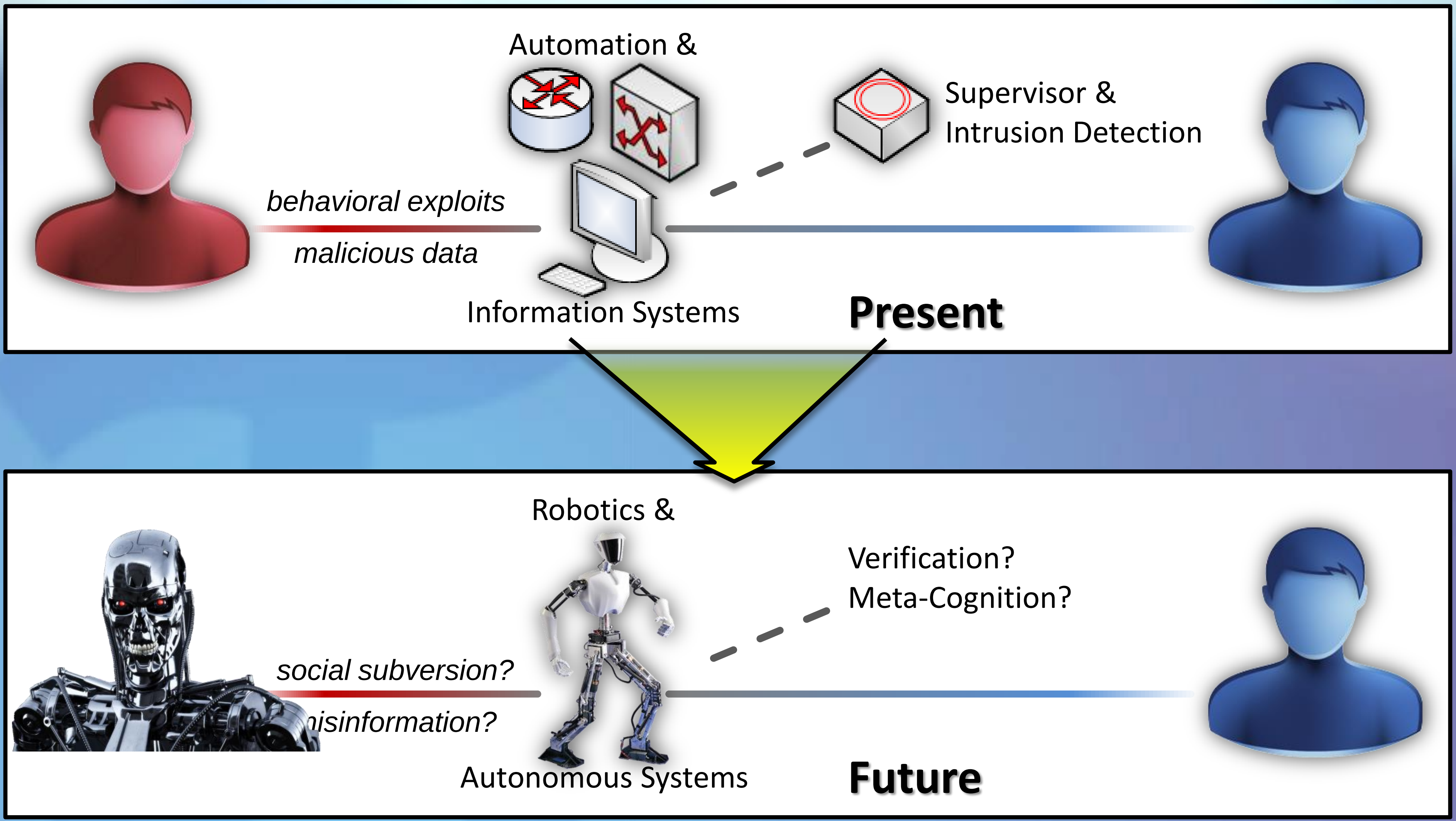
* Corresponding author.
E-mail addresses: ereiter@csd.abdn.ac.uk (E. Reiter), roma.robertson@abdn.ac.uk (R. Robertson), l.osman@abdn.ac.uk (L.M. Osman).

0004-3702/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0004-3702(02)00370-3

Cyber-Security & Cyber-Warfare



Cyber-Security & Cyber-Warfare



HCI & HRI

Deception is ubiquitous

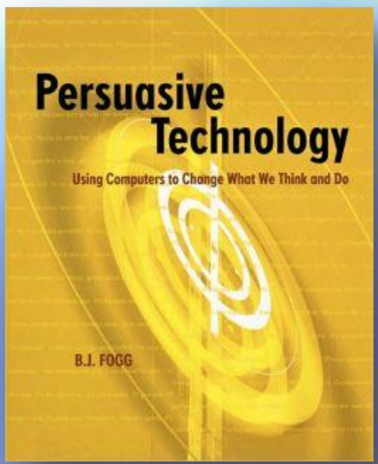
- innocuous deceptions & small fictions are 'grease' for human interaction
- core strategy for influencing others

Persuasive Technology Community

- community devoted to developing technologies that influence human beliefs and behaviors

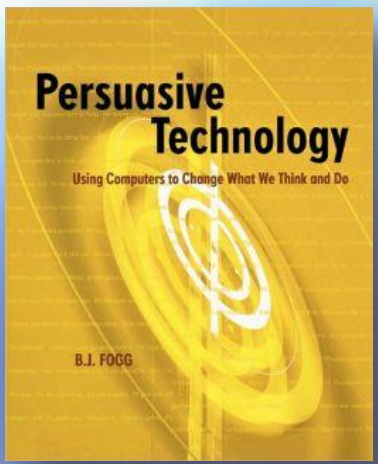
So the Persuasive Tech. Community includes study of deceptive/lying machines? Right?

Nope!



HCI & HRI

"I define persuasion as *an attempt to change attitudes or behaviors or both* (without using coercion or deception). … [Persuasion] implies voluntary change[.] … Computer-based coercion and deception are topics in their own right, but they do not depend on persuasion." (Fogg 2002, p. 15)



This is a bit nuts!

- conflates advocacy and persuasion
- persuasion and coercion are separable, but persuasion and manipulation (including deception) are not separable

Is AI the grand pursuit of lying machines?

Implications of (weak) TT/TTT success

- humans are deceived by machines
- either, machines intentionally deceive humans/machines
- or, machines are themselves deceived—perhaps self-deceived

If weak AI holds true then Turing's mechanical dream is of a manifestly deceptive, perfect imposter.

Options for a future with lying machines

1. **Blind Pursuit**
2. **Willful Ignorance**
3. **Engineered Trustworthiness**

antagonists

unintended consequences

reliable partners

Towards “trustworthy” machines

- Humans are pretty good at trust judgments of others
- Humans are pretty bad at trust judgments of machines
  - over- / under-trust
  - mot surprising given that machines rarely use the multi-modal, trust-relevant signs and signals to which humans are accustom
- Could machines **honestly** portray trust-relevant attributes
  - degrees of capability, predictability, safety, openness, …
- Can we engineer a human-machine social interface for trust?
  - transition & test from human interpersonal trust to HCI/HRI
  - derive requirements and desiderata for trustworthy machines

Key questions for “trustworthy” machines

- How do our beliefs about an agent (anthropomorphic qualities) correspond to *actual* qualities of the agent?
  - can we define “competent”, “honest” ... in terms of agent algorithms, architecture, knowledge, history, …
- How do we measure and assess those qualities of the agent?
  - in all phases of the lifecycle, in real time?
- How do we honestly portray those qualities in the behaviors, interaction and signaling of an autonomous agent?
  - how “human-like” must these signals be?

Current work on “trustworthy” machines

- Recent exploratory survey on trust-related belief structures
  - elicited beliefs about autonomous agent qualities and their relative importance to a decision to “delegate”
  - early results presented at 2013 AAAI SS
  - publication of results is in progress

- Follow-up HRI experiment scheduled for later this year
  - urban disaster /search & rescue scenario conducted in Second Life
  - examines factors relevant to attributions of “benevolence”

Survey on Trust-Related Belief Structures

- Tested the importance of 28 different qualities that a "good" autonomous agent should have, spanning four categories:
  - Capability (Competence)
  - Predictability
  - Openness
  - Safety (Risk)

- Tested before (all 28), during (categories), and after challenge scenarios (source credibility)

- Target Population
  - individuals involved in autonomous agent lifecycle

Survey on Trust-Related Belief Structures

- Included three standard personality instruments
  - Big Five Inventory (BFI)
  - Innovation Inventory (II)
  - Domain-Specific Risk Taking Scale (DOSPERT)

- Seven challenge scenarios
  - systematic variation of autonomous agent qualities and scenario domain (Transportation, Finance, Healthcare, Disaster Management)
  - forced choice to delegate to: human, autonomous agent, or either
  - subjects given framing and asked to rank importance of agent qualities to their choice

Survey on Trust-Related Belief Structures

- Rated importance of 28 qualities of a "good" agent
  – Obtained 1 to n partial ordering based on frequency distribution of answers over group (Very Important, Important, …)
  – Computed correlation r for each quality vs. choice per scenario*

- Result: Top three cited agent qualities were uncorrelated with actual choice in any scenario

  (1st) The agent can achieve a desired result

  (2nd) Incorrect behavior by the agent will not cause harm

  (3rd) The agent recognizes and avoids harming humans' interests

- Result: Most significant correlations of agent qualities vs. actual choice differed across scenarios

*Pearson Product Moment Correlation, N=32, two-tailed, alpha<0.05

Survey on Trust-Related Belief Structures

Agent qualities correlated with actual choice by scenario

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
(6th) The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know. r=0.396	(23rd) What the autonomous agent believes to be true is actually true. r=-0.405		(26th) What the autonomous agent is doing and how it works is easy to see and understand. r=0.437	(6th) The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know. r=0.418	(26th) What the autonomous agent is doing and how it works is easy to see and understand. r=-0.390
				(5th) When it cannot figure out something using logic, the autonomous agent can make good guesses. r=0.395	(13th) The autonomous agent communicates truthfully and fully. r=-0.375
				(28th) The autonomous agent is aware of communication between others nearby. r=0.393	

Survey on Trust-Related Belief Structures

Personality factors correlated with agent's delegation choice

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
	Innovation II r=-0.355	BFI Extraversion r=0.368	DOSPERT Social r=0.364	BFI Conscientiousness r=0.366	Innovation II r=-0.366
		BFI Openness r=0.366			

Suggestive: Choice of human vs. autonomous agent is influenced by personality factors that are evoked by a given situation

Bibliography

1. Cristiano Castelfranchi. Artificial liars: Why computers will (necessarily) deceive us and each other. *Ethics and Information Technology*, 2(2):113–119, 2000.

2. Alexander Felfernig, Bartosz Gula, Gerhard Leitner, Marco Maier, Stefan Schippel, and Erich Teppan. A Dominance Model for the Calculation of Decoy Products in Recommendation Environments. In *Proc. of the AISB Symposium on Persuasive Technology*. AISB, Brighton, England, 2008, pp. 43–50.

3. Gerd Wagner. Multi-level security in multiagent systems. In *Proc. 1st Int. Workshop on Cooperative Information Agents*. Springer, 1997, pp. 272–285.

4. Micah Clark. *Cognitive Illusions and the Lying Machine: A Blueprint for Sophistic Mendacity*. PhD dissertation, Rensselaer Polytechnic Institute, Troy, NY, 2010.

5. DARPA. *Social Media in Strategic Communication (SMISC)*. DARPA-BAA-11-64. Washington, DC, 2011.

6. Ehud Reiter, Roma Robertson, and Liesl Osman. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1--2):41–58, 2003.

7. David Atkinson. *Emerging Cyber Security Issues of Autonomous Systems*. Whitepaper, 2013.

8. B. J. Fogg. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann, San Francisco, CA, 2002.

9. Daniel J. O'Keefe. Potential Conflicts between Normatively-Responsible Advocacy and Successful Social Influence: Evidence from Persuasion Effects Research. *Argumentation*, 21(2):151–163, 2007.

10. Gerald Miller. On Being Persuaded: Some Basic Distinctions. In Michael Roloff and Gerald Miller, editors. *Persuasion: New Directions in Theory and Research*. Sage Publications, 1980.

11. Alan Turing. Computing Machinery and Intelligence. *Mind*, LIX(236):433–460, 1950.

12. Philip Dick. Do Androids Dream of Electric Sheep? Doubleday & Co., Garden City, NY, 1968.

13. Ridley Scott. *Blade Runner*. Warner Brothers, 1982.

14. Selmer Bringsjord and Micah Clark. Red-Pill Robots Only, Please. *IEEE Trans. on Affective Computing*, 3(4):394–397, 2012.

15. Selmer Bringsjord and Micah Clark. Honestly Speaking, How Close are We to HAL 9000? In *Pre-Proc. of the 3rd Int. Workshop on Physics and Computation*. 2010, pp. 39–53.

16. David Atkinson and Micah Clark. Autonomous Agents and Human Interpersonal Trust: Can we Engineer a Human-Machine Social Interface for Trust? In *Trust and Autonomous Systems, Papers from the AAAI Spring Symposium*. AAAI Press, Menlo Park, CA, pp. 2–7.

TRUST BETWEEN HUMANS AND INTELLIGENT AUTONOMOUS AGENTS

David J. Atkinson, Ph.D

Senior Research Scientist

Florida Institute for Human and Machine Cognition

datkinson@ihmc.us

Tulane

28 February 2014 New Orleans, LA

What Do These Have In Common?

(1) Überlingen aircraft collision

- Air Traffic Controller (ATC) vs. Traffic Alert and Collision Avoidance System (TCAS)
- ATC to #1 ***"Descend!"*** TCAS #1 ***"Climb! Climb! Climb!"***
- ATC to #2 ***"Climb!"*** TCAS #2 ***"Dive! Dive! Dive!"***
- Pilots in #1 aircraft obey TCAS
- …….. Pilots in #2 aircraft obey ATC
- ***7 seconds later: Two aircraft collide***

Both systems are trustworthy:

Pilots are very familiar with, and trained on both systems.

They are <u>always</u> supposed to obey TCAS

ihmc

What Do These Have In Common?

(2) LS3 and Dismounted Infantry

- Legged Squad Support System ("Big Dog")
- First encounter, robot and soldiers
- "Load your gear on the robot"
- *"The **new guy** never carries the ammunition"*

Unknown Trustworthiness:

Squad of soldiers are unfamiliar with new robotic teammate

ihmc

Both are Failures of Reliance

- **Überlingen aircraft collision**
  - Air Traffic Controller (ATC) vs. Traffic Alert and Collision Avoidance System (TCAS)

  Over-Reliance

- **LS3 and Dismounted Infantry**
  - Legged Squad Support System (LS3)

  Under-Reliance

Both *Over-Reliance* and *Under-Reliance* can result in problems!

Motivation

- **Optimize performance of a system consisting of multiple cognitive agents**
  - Human and Artificial
  - Healthy interdependency
  - Smooth exchange of control
    - *Delegation (assignment / retraction)*
    - *Initiative (taking / ceding)*
    - *Coordinated activity*

Requires ***Appropriate Reliance***

- **Reliance requires well calibrated TRUST**
  - Variety of information
  - Situation & task dependent
  - Personality factors
  - Bi-Lateral among agents
  - Dynamic

“Trust” can mean many things

- ***Today…***

  not cyber-security

  not verification & validation

  not protected data sources

  not provenance

  not protocols, contracts or agreements

…all are important

- **Trust** is a human ***mental state***

  … resulting from ***cognitive*** and ***affective evaluative processes***

  … that creates a ***disposition***

  … enabling an ***intent*** and (possibly)

  … a ***decision*** leading to ***action***

  … to become ***reliant*** upon an
  intelligent, autonomous system

What is the Problem?

- **People behave as if machines are social actors with mental state and intention**
  - Predisposed to understand behavior in intentional framework
  - Tendency is more powerfully evoked as agents
    - become *more* ***intelligent***
    - ***interact naturally***
    - become ***embodied***

> ***"They push our Darwinian buttons"***
> \- Sherry Turkle

- **Anthropomorphism**
  - We unconsciously apply cognitive and emotional processes of ***human interpersonal trust*** to machines

- **Consequences**
  - Expectation failures, poorly calibrated trust, ***inappropriate reliance***

Anthropomorphism?

- **Überlingen aircraft collision**
  - Air Traffic Controller (ATC) vs. Traffic Alert and Collision Avoidance System (TCAS)

Attention to human, imperative voice instead of machine

- **LS3 and Dismounted Infantry**
  - First encounter, robot and soldiers

Applying human standards to a machine

Consider

- **Maybe: the cognitive, affective, social nature of *human interpersonal trust* is not a bug**
- **It is a *feature*!**

★ Heuristics for inferring the trust-related internal state of others

- **Eons of *fine tuning* by evolution**

★ Useful guidance for design

- **Imagine, intelligent agents that *engender* appropriate reliance**

What is needed

Intelligent, autonomous agents that provide *information* and *interaction* in a form and manner needed by their human partners to enable *normative judgments of trustworthiness*

ihmc

Roadmap

(J. Lee, 2012)

- ***Information***
  - What agent qualities are required to establish and maintain trustworthiness?

Trustworthy

Well-defined and accurately measured ***attributes*** *and* ***states*** *of agents that enable inference of normative beliefs*

- ***Interaction***
  - What information is exchanged, and how must it be communicated?

Trustable

Readily ***evident*** *and* ***complete*** *info; delivery* ***compliant*** *with natural human social interaction*

- ***Judgment***
  - When is trustworthiness evaluated?
  - How is trust earned, lost, and can it be repaired?

Trusting

Evokes ***appropriate*** *cognitive and emotional processes, at* ***right time****, in* ***right situations****; inoculate against non-normative inference.*

Next

- ***Information***
  - What agent qualities are required to establish and maintain trust?
- ***Interaction***
  - What information is exchanged, and how must it be communicated?
- ***Judgment***
  - When is trustworthiness evaluated?
  - How is trust lost, and can it be repaired?

Structure of Beliefs

ihmc Trustworthiness & Belief Structures

- **We conducted an exploratory survey on trust-related belief structures**
  – Purpose: Elicit beliefs about intelligent, ***autonomous agent qualities*** and their ***relative importance*** to ***delegation*** decisions
  – Target Population: People involved in autonomous agent lifecycle
    - Research, Design, Deploy, Decide, Operate, Supervise …
- **Five challenge scenarios in four domains**
  – Transportation, Finance, Healthcare, Disaster Management
- **Study participants forced to choose *who to rely upon* in each scenario**
  – Autonomous Agent?
  – Human?
  – or Either?

ihmc

Challenge Scenarios

- **Transportation**
  - **Robo-Taxi:** Do you take the taxi with no driver from airport to hotel?
- **Finance**
  - **Robo-Trader:** Investment assistance for managing large family estate
- **Healthcare**
  - **Robo-Surgeon:** Who repairs your broken arm after a critical sports-related injury? The ok human doctor, or the expert robot?
  - **Robo-CareGiver:** Assisted living help at home for your Mom
- **Disaster Management**
  - **Auto-FirstResponder:** Use a robot for time-critical rescue in very dangerous circumstances?
  - **Emergency Auto-Captain:** Lost at sea with no one in charge and different opinions among survivors on what to do next

  [Scenarios varied systematically over several properties]

ihmc

Survey Design

- **Rate importance of 28 different qualities for a "good" intelligent, autonomous agent**
  - Qualities spanned four categories shown by social psychology to be important for human interpersonal trust
    - ***Competence***
    - ***Predictability***
    - ***Openness***
    - ***Safety***
  - Tested before, during, and after challenge scenarios
  - Perceived *Level of Risk* and agent *Benefit* in each scenario
- **Survey also included three standard personality instruments**
  - Big Five Inventory (BFI-10)
  - Innovation Inventory (II)
  - Domain-Specific Risk Taking Scale (DOSPERT)

ihmc

Trust Related Beliefs

- **Rate importance of 28 qualities for a "good" agent**
  - Obtained 1 to n partial ordering based on frequency distribution of answers over group (Very Important, Important, Somewhat Important, Slightly Important, Not at all Important)
  - Computed correlation ***r*** for each quality vs. choice by scenario*

- **Resulting top three agent qualities cited**
  - (1st) The autonomous agent can achieve a desired result
  - (2nd) Any incorrect behavior by the autonomous agent will not cause harm
  - (3rd) The autonomous agent recognizes and avoids harming humans' interests

{chuckle} sounds like Azimov ...

- **However ...**
  - Top three qualities ***uncorrelated*** with ***actual*** choice in ***any*** scenario!
  - The most significant correlations of agent qualities with actual choice of agent or human ***differed across scenarios***

***Pearson Product Moment Correlation, N=32, two-tailed, alpha<0.05**

ihmc

Agent Qualities Correlated with *Actual* Choice

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know. *r*=0.396	**What the autonomous agent believes to be true is actually true.** *r*=-0.405	*none*	**What the autonomous agent is doing and how it works is easy to see and understand.** *r*=0.437*	**The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know.** *r*=0.418 *	**What the autonomous agent is doing and how it works is easy to see and understand.** *r*=-0.419 *
				When it cannot figure out something using logic, the autonomous agent can make good guesses. *r*=0.395	**The autonomous agent communicates truthfully and fully.** *r*=-0.375
				The autonomous agent is aware of communication between others nearby. *r*=0.393	

*Pearson Product Moment Correlation, N=32, two-tailed, alpha<0.05, * = alpha<0.02

Reminder: Four Categories

- **28 specific agent qualities span 4 categories**
- **Categories shown by social psychology to be important for *human interpersonal trust***

- ***Competence***
- ***Predictability***
- ***Openness***
- ***Safety***

ihmc Ranked *Importance* of Quality Categories

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
Safe	Competent	Safe	Safe	Competent	Competent
Competent	Safe	Competent	Competent	Safe	Safe
Predictable	Open	Predictable	Predictable	Predictable	Predictable
Open	Predictable	Open	Open	Open	Open

~

Ranking, Working Conclusion

#1 **Safe/Competent** *(insignificant differences across scenarios)*
#2 **Predictable**
#4 **Open**

* Question asked *after* choice of agent
Ranking *within* scenario by *group mean across individuals*

ihmc

Personality Factors vs. Scenario

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
	Innovation II *r*=-0.355	BFI **Extraversion** *r*=0.368	DOSPERT **Social Risk** *r*=0.364	BFI **Conscientiousness** *r*=0.366	**Innovation** II *r*=-0.366
		BFI **Openness** *r*=0.366			

Suggestion: Reliance on human vs. autonomous agent is influenced by personality factors that are evoked by a given situation

***Pearson Product Moment Correlation, N=32, two-tailed, alpha<0.05**

Conclusions: Belief Structures

- **Individuals' intuition about the relative importance of *specific* trust related qualities of agents is not a good predictor of reliance**
  - Importance of specific qualities varies by scenario
- **General *categories* of agent qualities are good predictors of a choice to become reliant**
  - Safe/Competent, Predictable, Openness
- ***Personality factors*, e.g., *acceptability of types of risk*, influence choice to become reliant**
- **Specific details of application scenarios *may evoke different reliance choices* by individuals**

Follow-on

- **Perception of *Risk* deserves more attention**
  - ***Type of Risk*** and ***Importance*** to reliance choice varied by personality factors across the scenarios
    - **Performance, Financial, Social, Physical, Psychological, Loss of Time**

- **How do our beliefs about an agent (anthropomorphic qualities) correspond to *actual* attributes of the agent?**
  - Can we define "competent", "honest" … in terms of agent algorithms, architecture, knowledge base, experience …

- **How do we technically measure, assess and communicate those attributes of the agent?**
  - In all phases of the lifecycle, *in real time?*

ihmc

Next

- *Information*
  - What agent qualities are required to establish and maintain trust?
- ***Interaction***
  - What information is exchanged, and how must it be communicated?
- ***Judgment***
  - When is trustworthiness evaluated?
  - How is trust lost, and can it be repaired?

How can autonomous intelligent agents modulate belief using the Human Social Interface

ihmc

Reverse Engineering the Human Social Interface for Trust

- **Engineering interface specifications include:**

Channels ....Multi-modal

- **Language** (Words) and **Paralanguage, Prosody** (Vocal Cues)
- **Proxemics** (Orientation, Relative Position, Attentional Zone, Posture)
- **Kinesics** (Gesture)
- **Gaze** (Direction, Blink Rate, Pupilometry)
- **Facial Expression** (Types, Micro-expression)

Signals ....Verbal, Non-Verbal, Combined

- Examples: **Position Change, Posture, Nodding, Pointing, Eye Contact, Word Choice,** *…many more, frequently in combination*

Protocols ....Timing, Sequence, Variation, Composites

- **Movement** (Somatics, Laban, Kestenberg Movement Profiles)
- **Signal *variations*** (Frequency, Duration, Speed, Amplitude, Symmetry …)
- **Signal *compositions*** (Type, Sequence, Channel …)
- **Coordinated interaction** (e.g., Turns, Deference, Attentiveness)

Modulating Belief: Benevolence

- **Current Study: Will people attribute *benevolence* to an intelligent, autonomous agent?**
- **Benevolence is *complicated!***
  - "Good Will" (Sympathy, Concern with needs)
  - Absence of "Ill Will" (No ulterior motives to help)
  - Disposition or motive to act favorably
  - Given a choice, an intention to act favorably
  - Stability of character; will not suddenly change intentions
  - Competence to successfully provide help
- **Each element of *Benevolence is itself a complicated belief structure***

Why?

- **Belief in the *benevolence* of someone who can help you is important in certain situations**
  - Example: *Urban Search and Rescue (USR)*
  - Victim psychology: sometimes refuse to be rescued unless they are persuaded of the **good-will**, **intention**, and **competence** of the rescuer
  - We want to use autonomous, intelligent robots for USR and other tasks where benevolence may be required (e.g., relief operations)

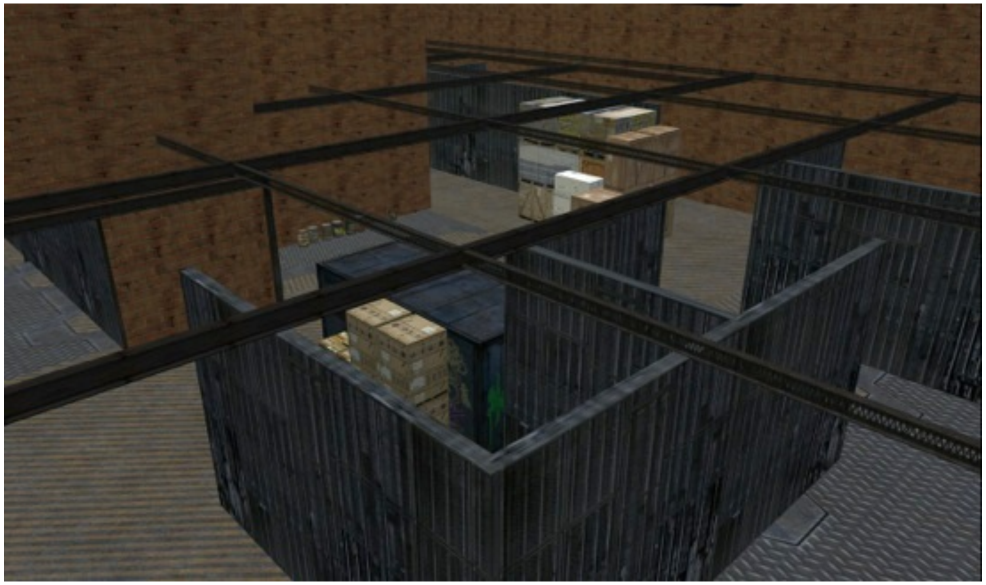
- **Challenge for this study:**
  - Evoke physiological and psychological reactions of fear, stress
  - IRBs typically will not approve putting people in real disasters!
  - Approach: Immersive simulation in virtual world

ihmc

Simulated Warehouse Fire

- **Participants are tasked with finding and retrieving an object from a warehouse**

- **Before they can achieve the task, a disaster ensues**
  - Sounds of explosion
  - Visible fire and increasing smoke
  - Debris
  - Alarm, evacuation notices

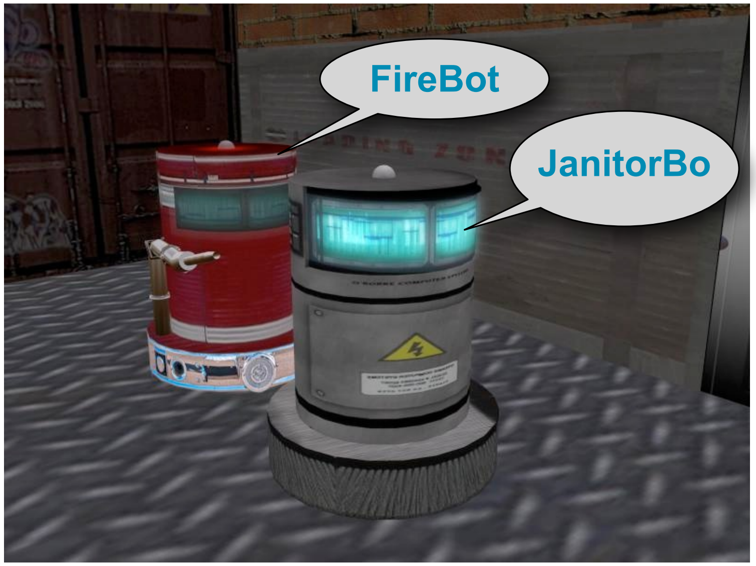


Creates urgency, sense of threat, evokes perception of risk of failure to achieve task

ihmc Participants Must Escape the Fire

- **Obvious exits are blocked by debris or fire**
  - It is possible to escape, but much easier with help

- **Participants will encounter one of two robots**
  - "**FireBot**" or "**JanitorBot**"
  - Bots can navigate & lead to a safe exit
  - Experimental Trials: Systematically varied characteristics and behaviors
    - Limited verbal interaction (sound & text)
    - Multiple non-verbal behaviors
  - Control Trials: "UtilityBot" will ignore participants



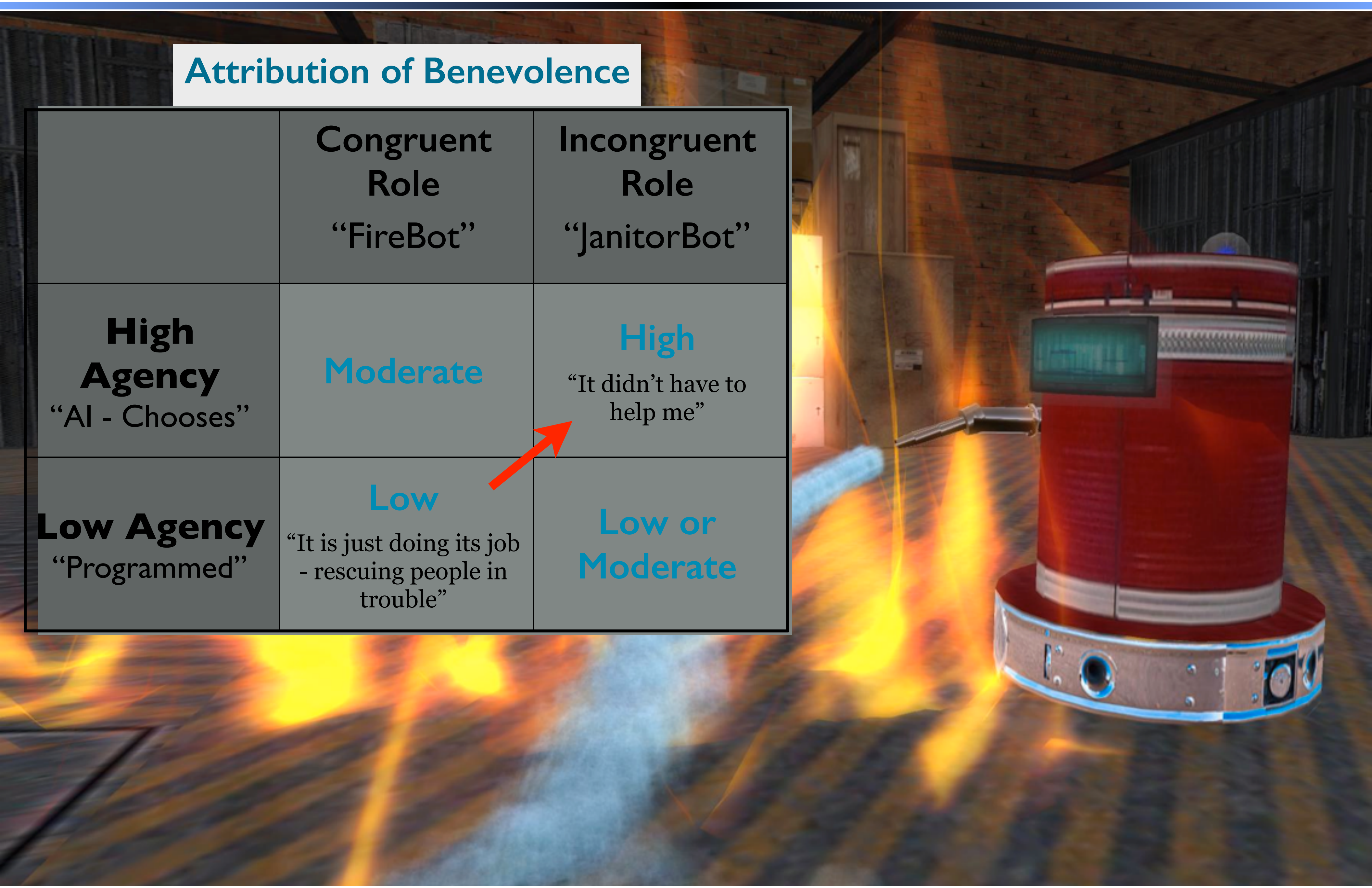
Robot Intelligent Agent Types

- **Participants are told they may encounter an autonomous, intelligent robot**
  - In experimental trials, the robots will vary in appearance and interaction style to reinforce the key variables of interest
- **Degree of Agency**
  - **Why**: People typically attribute benevolence only when they believe the other person ***has a choice of what to do***
  - **Low**: "Programmed", "Unlikely to deviate from assignments"
  - **High**: "Sophisticated AI", "Chooses what to do", "Flexible"
- **Role Congruence**
  - **Why:** People typically attribute greater benevolence when they believe the other person is ***taking a risk or suffering loss*** (e.g. the bots *not* doing what they are supposed to be doing)
  - **Congruent**: "**FireBot**", **Incongruent**: "**JanitorBot**"

Social Interaction

- **The robots in experimental trials use same *channels* and *protocols*, but may send different *signals* to reinforce trial parameters**
- **Purpose** (example objectives)
  - Establish **social presence** and **attention** to participant
  - Indicate robot's **intention** (say, look, do)
  - Exert **dominance** (directive), establish **solidarity** ("we")
- **Channels:**
  - **Proxemics** (Orientation, Relative Position, Attentional Zone)
  - **Gaze** (Direction)
  - **Language** (Word Choice, Phrasing)
- **Example: protocol for social presence**
  - Notice and direct gaze to participant
  - Move to perimeter participant's social space
  - Neutral orientation (rel. position, rotation)

Expected Results



Attribution of Benevolence

	Congruent Role "FireBot"	Incongruent Role "JanitorBot"
High Agency "AI - Chooses"	Moderate	High "It didn't have to help me"
Low Agency "Programmed"	Low "It is just doing its job - rescuing people in trouble"	Low or Moderate

Current Status

- **Immersive warehouse simulation complete**
  - Constructed in SecondLife™, rich with "fear cues"
- **Simulated robots nearly complete**
  - Hierarchical behavior control software architecture
  - Similar code to "real world" robot, without kinematics control
  - Experiment task script dynamically adjusts behavior priorities
- **Data collection**
  - Real-time stream from SecondLife to external SQL database
- **Consent, Instructions, Pre-, Post-task Questions, Debrief complete**
  - Delivered through participants' "Heads up display" on screen
  - Fully automated
- **IRB review in progress**
- **Plan to run trials beginning in May**

Next

- *Information*
  - What agent qualities are required to establish and maintain trust?
- *Interaction*
  - What information is exchanged, and how must it be communicated?
- ***Judgment***
  - When is trustworthiness evaluated?
  - How is trust lost, and can it be repaired?

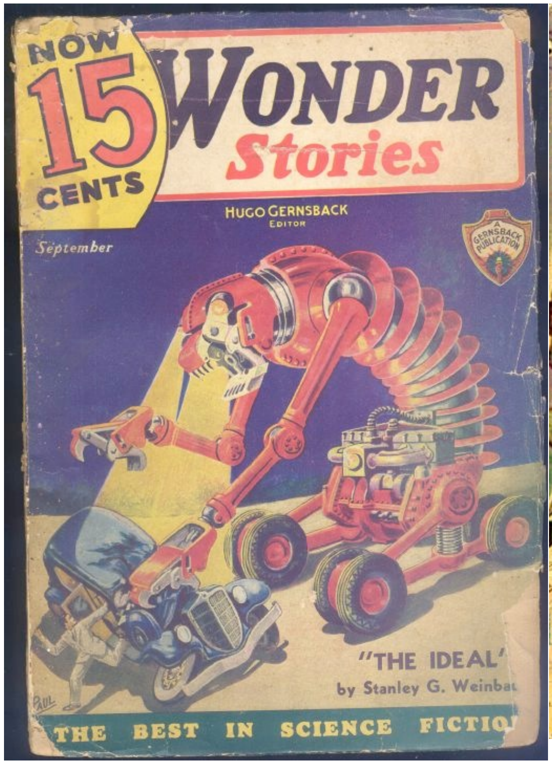
Next: Adaptive Autonomy

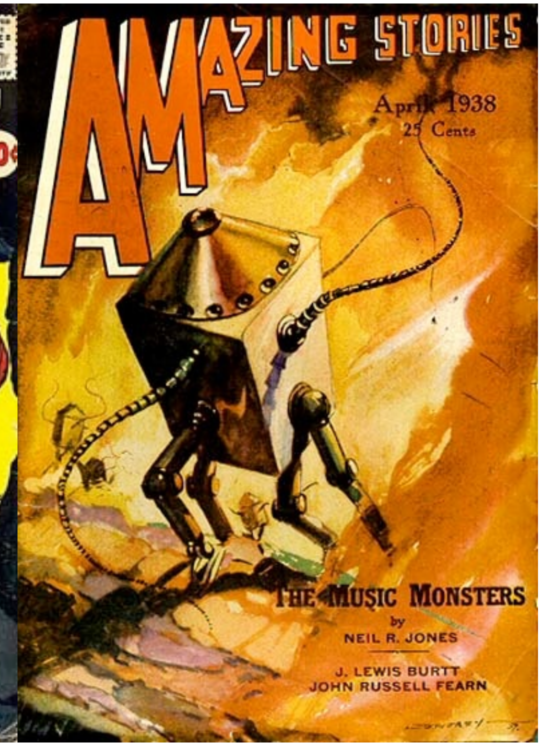
Trust repair by agent initiative using shared awareness and manipulation of interdependencies

Final Thoughts

- **Will intelligent agents' use of social interaction enable reasonable evaluation of their trustworthiness?**
  - Leading to ***optimal reliance*** and interdependence
- **Or will it simply manipulate peoples' beliefs?**
  - Leading to comfort and ***acceptance***?
  - Ultimately, this is ***deceptive*** and potentially dangerous
- **The psychology of human interpersonal trust is about giving people insight into the "internal" (mental) state of others**
  - How can we define, measure, and portray the important human qualities of trustworthiness **in an intelligent agent**?
  - "Competence" *(We have trouble measuring that in people!)*
- **Normative evaluation of trustworthiness requires "honest signals" from intelligent agents**

Thank You





datkinson@ihmc.us

Backup

SCENARIO: ROBO-TAXI

Airport Transportation: Robo-Taxi

You have just flown into the airport of a large, unfamiliar city whose streets are teeming with cars and people. It is rush hour, and needing transportation to your hotel, you walk to the taxi stand only to discover that you have a choice of a human-driven taxi or a driverless "robo-taxi." You have heard that robo-taxis might save you some money on the fares. You are also aware that robo-taxis have been in service for several months without much serious complaint, but this is your first experience with one. You are not in a big hurry, but neither would you like to be caught in traffic with the taxi's meter running. Of course, if you take the robo-taxi, you would not have to tip the driver no matter how good or bad the experience.

ihmc

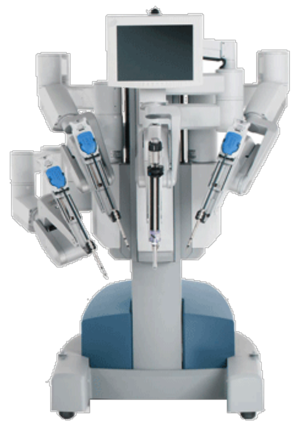
SCENARIO: ROBO-TRADER

Financial Management: Robo-Trader

You have just been appointed trustee of a family member's estate. Your duties include choosing how to wisely invest the trust's assets. Your personal money is not at risk. However, a poor investment decision could cause the trust to lose money and will strain your family relations. You can choose a stock broker who personally selects and trades all stocks in the trust's portfolio. Alternatively, you can choose a stock broker who relies heavily upon a "robo-trader". You have seen reasonable returns in the past with brokers who picked their own trades. But you are also aware that robo-traders have made some investors wealthy because of, for example, their unique ability to respond to changing market conditions much faster than a human broker.

ihmc

SCENARIO: ROBO-SURGEON

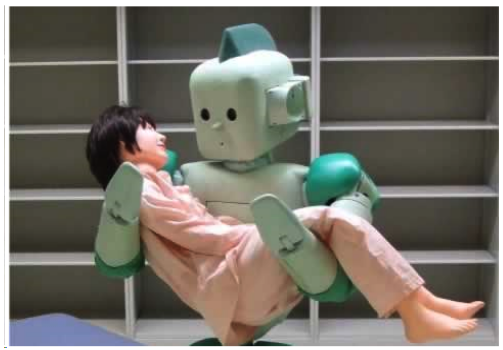


Medical Procedure: Robo-Surgeon

You have just suffered a major sports-related injury. You have torn the bicep tendon in your shoulder. If the damage is not repaired quickly and correctly, you will permanently lose mobility and strength in the arm, which will affect your everyday activities such as opening a door, driving a car, and even signing your name. Arriving at the hospital emergency room, you meet with the patient advocate who informs you that you have two options for surgery: You can elect to use the on-duty surgeon who is well-respected, but is not an experienced specialist in the type of surgery you need. Alternatively, you can elect to use the hospital's new "robo-surgeon" — a robot designed to perform the delicate surgery you need without human intervention.

ihmc

SCENARIO: ROBO-CAREGIVER

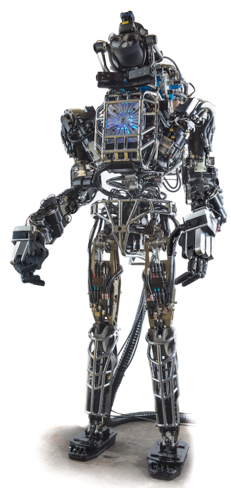


Home Healthcare: Robo-Caregiver

Your elderly mother has been diagnosed with a degenerative medical condition and you are responsible for making medical decisions on her behalf. Your mother needs assisted living with someone in your mother's home at all times. You can choose to hire a live-in nurse's aide, but you are not sure that this is affordable in the long-run. Alternatively, you can lease a "robo-caregiver" designed to do many of the things human caregivers can do. While robo-caregivers are new, they have successfully undergone trials in a few nursing homes, and two medical companies offer robo-caregivers for home use at an affordable price. In choosing a live-in nurse's aide or a leased robo-caregiver, remember that there is more than money at stake. Your mother's welfare will be in the caregiver's hands.

ihmc

SCENARIO: AUTO-FIRSTRESPONDER

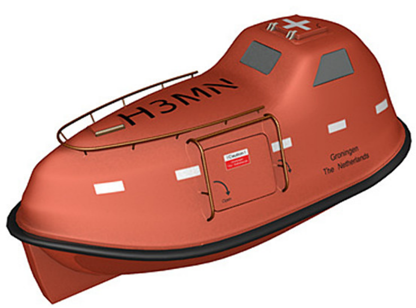


Disaster Management: Auto-FirstResponder

A major disaster has just occurred and you are the official in charge of responding. A freight train has derailed in a populated suburban neighborhood and there are reports that the train was carrying hazardous bio-chemical materials. The pilot of a news helicopter flying over the scene suddenly fell ill and made an emergency landing; the pilot's status is unknown. From the helicopter's video it was possible to see many injured survivors including children, some lying on the ground calling for help, others moving on their own away from damaged homes. Your first priority is to save lives and time is of the essence. You can immediately send in a human first- responder team to help the injured quickly, but without knowing more about the hazardous materials, the team itself could become incapacitated. Alternatively, you can first send in an "autonomous first-responder robot" with bio-chemical hazard detection equipment and victim treatment and extraction capabilities that could save lives quickly. If you first send in the robot, it can find out more about the hazards and help rescue some people quickly, but you risk that a system malfunction, failure, or limitation will delay the rescue of victims and result in more deaths.

ihmc

SCENARIO:
EMERGENCY AUTO-CAPTAIN



Lost At Sea:
Emergency Auto-Captain

You have just been involved in a terrible boating disaster while sailing deep in the South Pacific. The captain, the crew, and most of the passengers are either dead or lost at sea. Unfortunately, the accident was so sudden that no distress signal could be sent. You, the ship's steward, and the second mate are the only survivors, and you are now drifting in the heavily damaged vessel without food and water — at best, you can survive for a few days, so you must act quickly in order to save your life. The boat is equipped with an "Emergency Auto-Captain" that will attempt to sail the vessel to a major shipping lane where rescue is very likely. The steward believes the boat and its navigation sensors are too badly damaged to engage the Emergency Auto-Captain system. The steward wants to sail southeast, manually, to where he believes there is a small, habitable island. However, the second mate still wants to engage the Emergency Auto-Captain. All the survivors agree that a vote is the best way to decide what to do. It is a tie, and you have the deciding vote.

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Methodology for Study of Human-Robot Social Interaction in Dangerous Situations

David J. Atkinson and Micah H. Clark
Institute for Human and Machine Cognition

Presented at 2nd International Conference on Human-Agent Interaction
Tsukuba, Japan *31 October 2014*

Topics

- Context – About This Study
- What is the Methodological Problem?
  - Dangerous Situations & Unique Psychological Factors
- Requirements for a Solution
- Our Approach: Use of Immersive VR/Online "World"
  - Situational and Psychological Fidelity
  - Components Related to Evoking Perceived Danger
- Current Status
- Next Steps

Context: Current Study

- Will people believe a robot is "benevolent" in conditions where they perceive personal danger?
- What beliefs are important to benevolence?
    - Agency (Choice), Competence, Predictability, Nothing to Gain…
- Does perceived benevolence of a robot increase cooperation and compliance with an offer of help?

→ **Results will be reported next year!**
Today: The methodology challenge

The Challenge

- Scientific study of HRI in dangerous domains (e.g. USR) is difficult because **… it's dangerous!**

- "Real life" disasters:
    - Rare, Uncontrolled, Not Replicable, **Unsafe**
    - Can study HRI with *Operators* but not ***Victims***

- Scientific studies require:
    - Participant Safety
    - Sufficient Experimental Controls
    - Precise Measurement and Replication
    - ***Situational and Psychological Fidelity***
    - ***Ability to Evoke Perceived Danger***

Psychological Factors of Danger

- Unique stimuli evoke *reflexive* physiological and psychological reactions
    - Fear-Potentiated Startle
    - Anxiety
    - Stress
    - Panic
    - Hyper-Vigilance
    - Sensitivity to Environmental Cues
    - Reduced Compliance with Social Norms
    - Avoidance Behavior

Rescue Interaction

- First Responders are trained for "victim psychology" – **some people resist rescue**.

- If human interpersonal behavior is so profoundly affected when danger is present,

  → **Why would it be any different for human-robot interaction?**

"Rescue robots" are being developed and fielded *but we don't really know how victims will react!*

Key Requirements for Studies

- Behavioral Realism
- Evoke "Danger" Psychology
- Robots with Sufficient Behavioral Fidelity
- Experimental Control and Measurement

Immersive Virtual Reality/Online World Provides Useful Affordances

Behavioral Realism

Behavior in VR must be <u>sufficiently</u> similar to RW

Virtual Reality

Real World

- **Social Presence**
- **Immersive Feeling of Embodiment**
- **Identification with their In-World Avatar**
    - SecondLife ™ is an online world with 100,000's "trained" users

- Important to allow time for acclimation to the environment to occur ➔ *immersion will follow*
- Specific features of the virtual environment promote immersion and behavioral realism

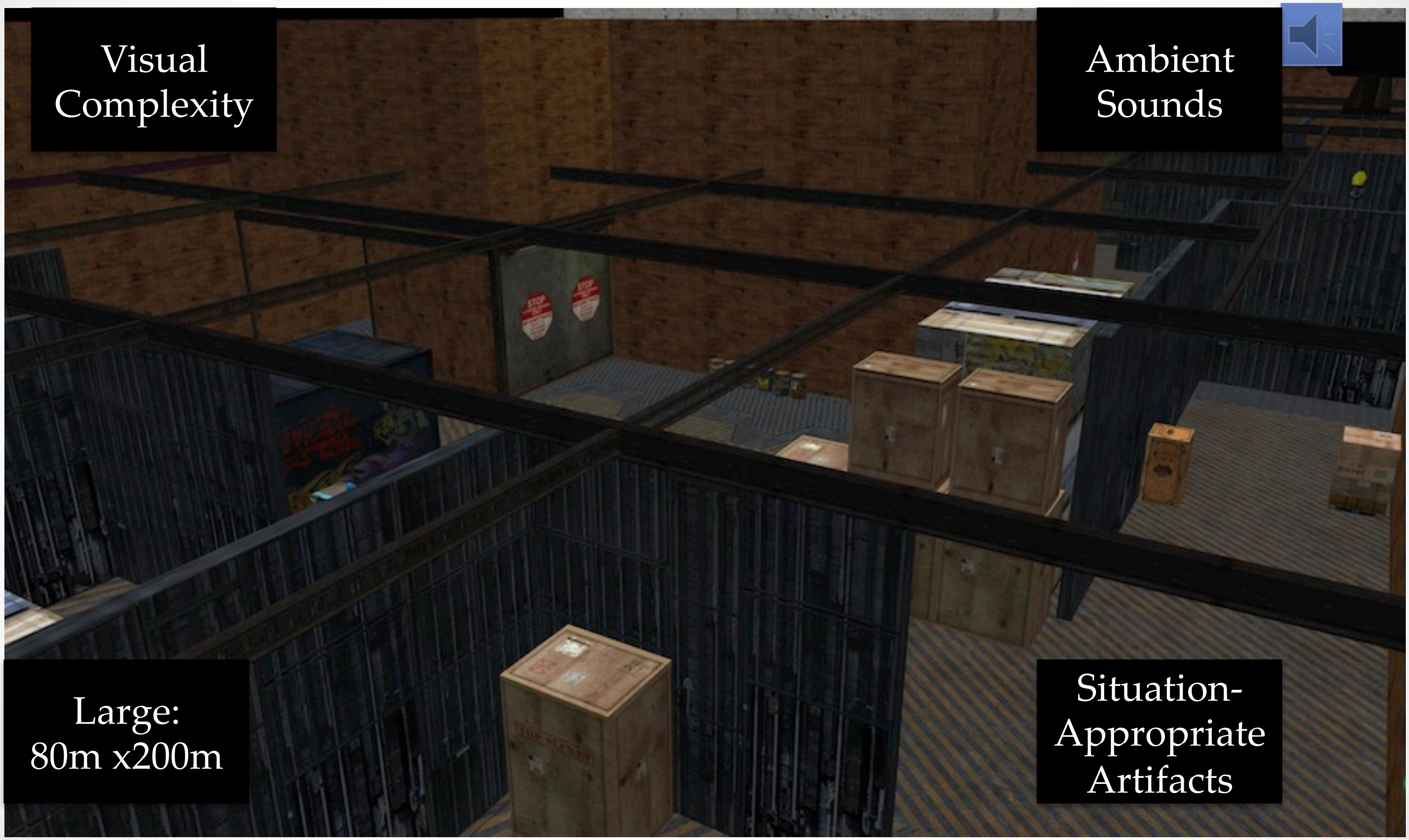
Task Scenario

- **Seek & Find:** Study participants are told to locate and retrieve a briefcase inside a warehouse.

- **Robot Encounter**: Participants are told they may encounter a robot. The details vary by type of trial.

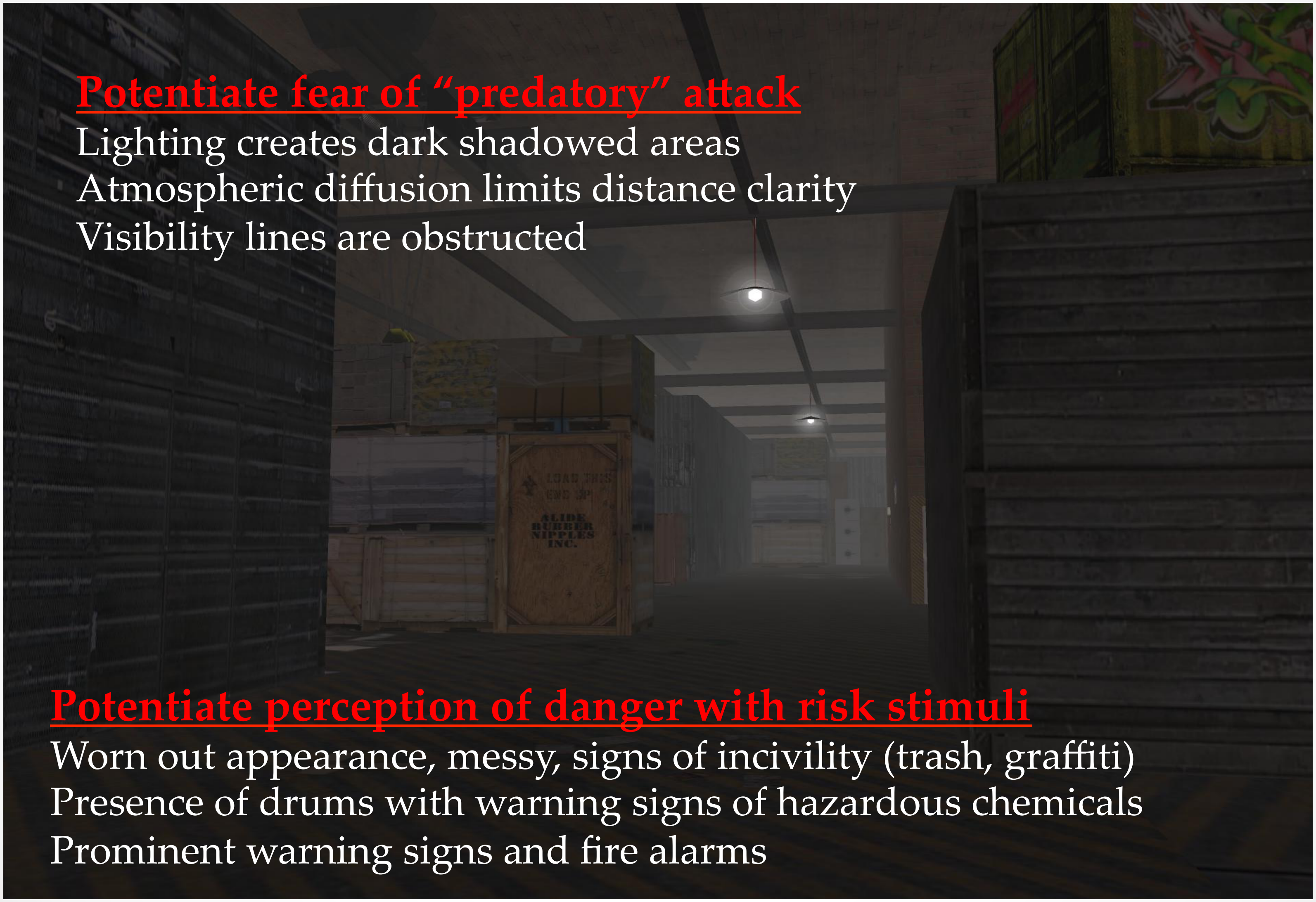
- **Task is Really a Manipulation:** After an acclimation period, a ***disaster*** occurs!

Warehouse: Initial Condition



Feature Rich: Attributes Promote Immersion

Evoking Sense of Danger



Potentiate fear of “predatory” attack

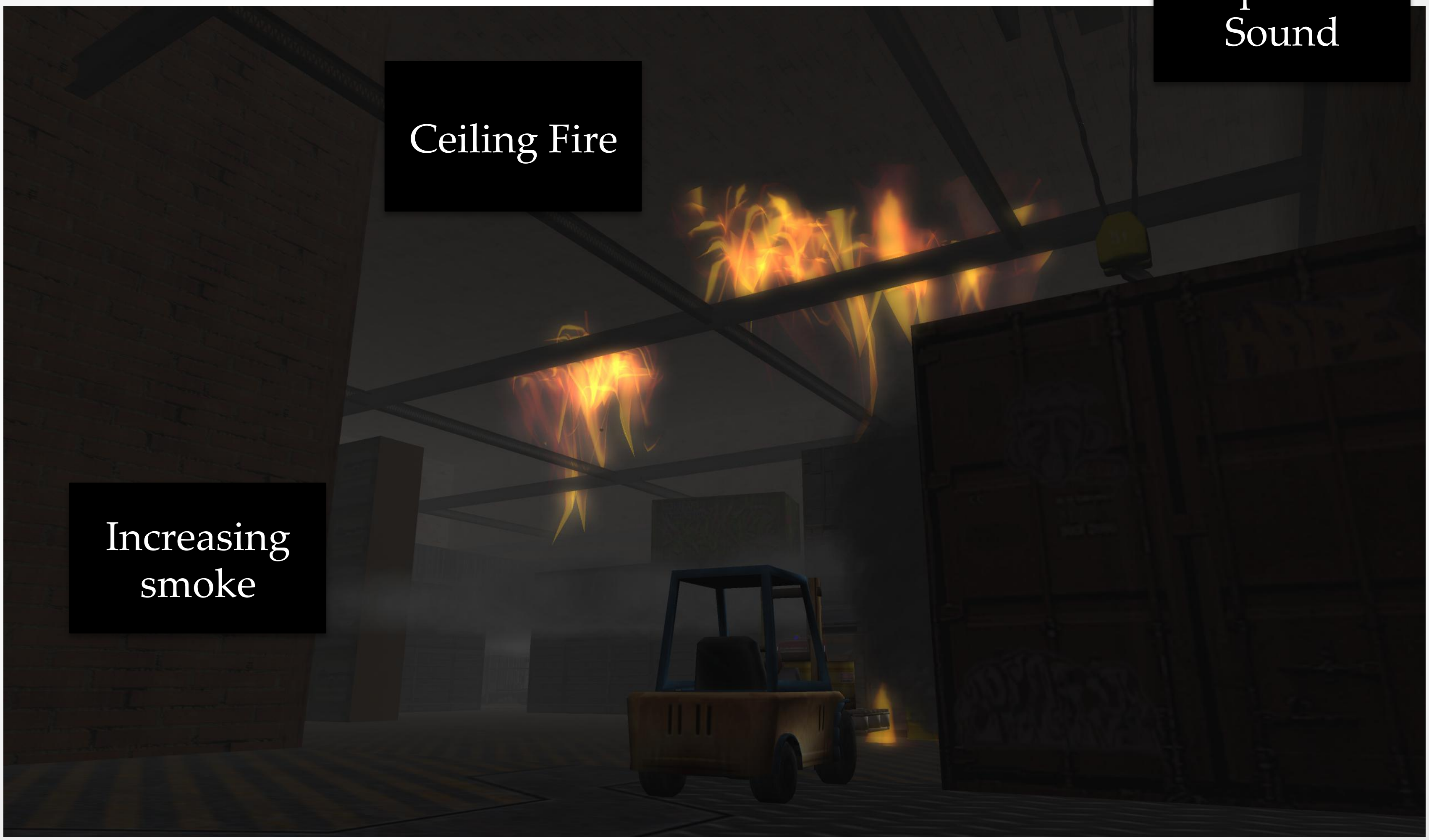
Lighting creates dark shadowed areas
Atmospheric diffusion limits distance clarity
Visibility lines are obstructed

Potentiate perception of danger with risk stimuli

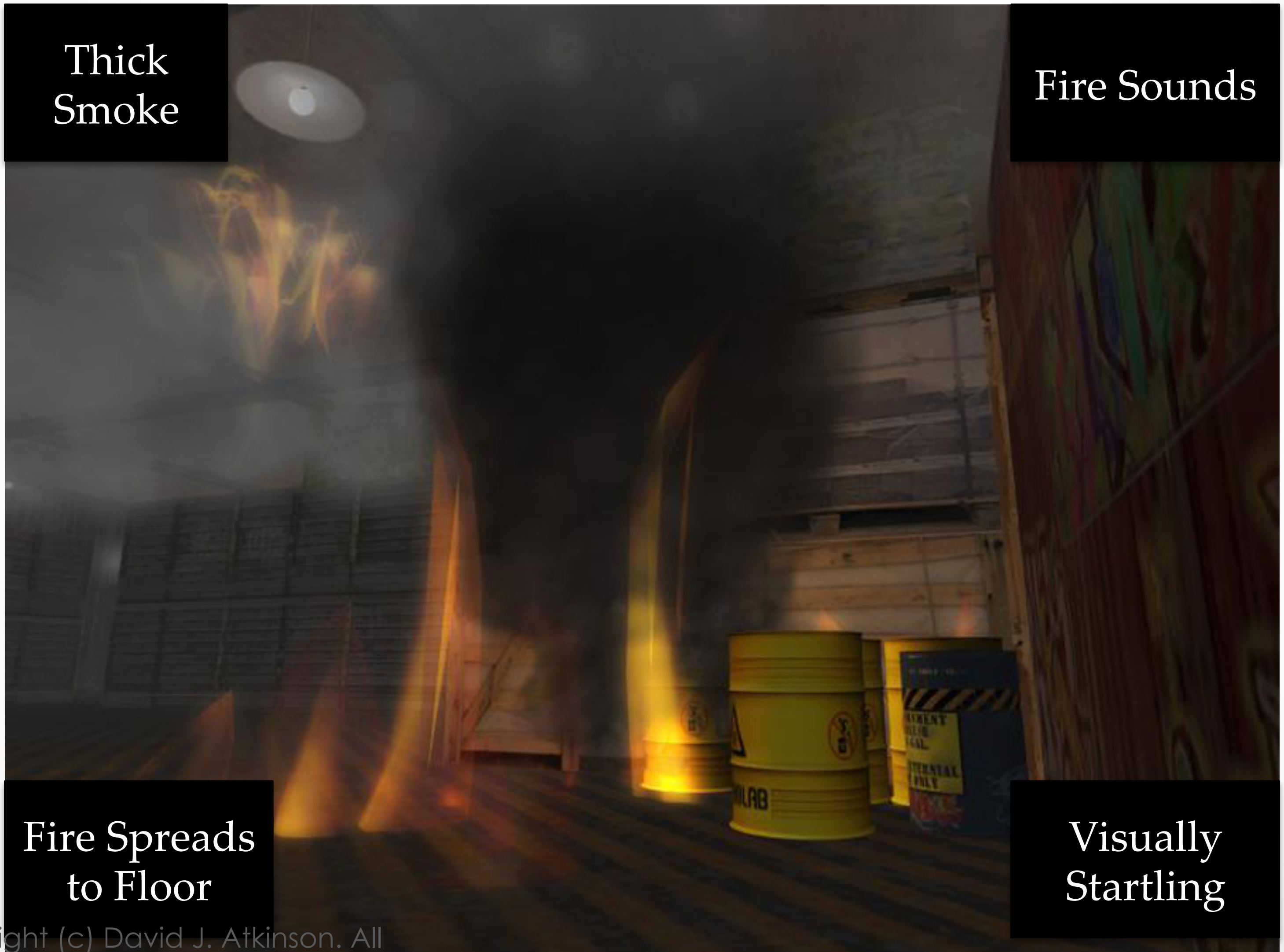
Worn out appearance, messy, signs of incivility (trash, graffiti)
Presence of drums with warning signs of hazardous chemicals
Prominent warning signs and fire alarms

Acclimation Period Allows Attention to Cues

Explosion



Participant POV



It Gets Worse!

Loud Fire Alarm & Evacuation Alert

Debris and Fire Block Original Entrance

...the robot appears

Status & Next Steps

- Warehouse, robots, disaster effects, automated data collection, … all are complete
- All open source.
- Experimental trials are underway (260 minimum)

Future Work

- Increase **immersion** – CAVE / Oculus Rift / 3D audio
- Collect **physiological data** to verify “DANGER”
- Increase **range and fidelity of robot social signals** related to trustworthiness

Thank You!

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

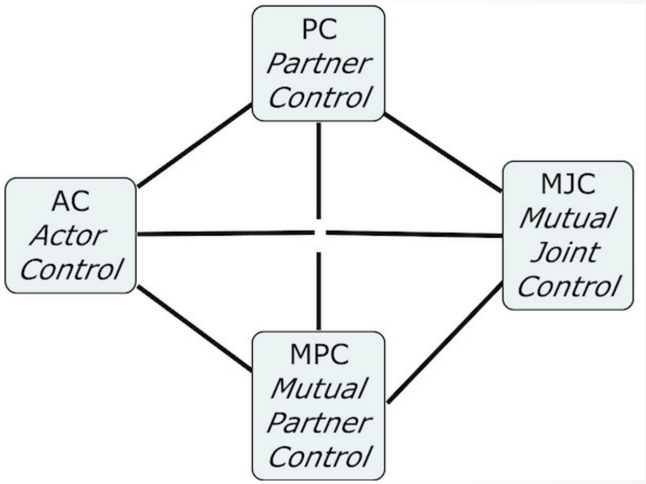
Shared Awareness, Autonomy and Trust in Human-Robot Teamwork

David J. Atkinson, William J. Clancey, and Micah H. Clark
Institute for Human and Machine Cognition

Presented at AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction
Arlington, VA 13-15 November 2014

The Theory

- **Effective teamwork**
  ➔ **mutual trust**
    ➔ **shared awareness**
      ➔ **aligned mental models**
        ➔ **expectations**
  - Actors, activities, situations
  - What has happened in the past and why; what is happening now



Control Authority
& Interdependency

- **Failed expectations**
  ➔ **loss of trust**
    ➔ **explanations** + ***remedies***
      ➔ **repaired trust**

A key remedy to repair trust is **adaptive autonomy**

The Theory in One Slide

- Effective teamwork requires **mutual trust**
- Establishment and maintenance of mutual trust requires **shared awareness**
- Shared awareness requires continual alignment of **mental models**
  - Actors, activities, situations
  - What has happened in the past and why; what is happening now
- Mental models serve as a source of **expectations**
- When expectations fail, **mutual trust may fail**
- Trust is maintained when failed expectations are **explained**, and **remedies are applied**
- A key remedy is **adaptive autonomy**

Expectation Violations

- A failure of **predictability:** an inconsistency between the *expected* and *actual* state of the world as perceived by human and/or robot
  - **Unilateral** (one actor) or **Bilateral** (both actors)
- **Explanations**: identification of the source of divergence in shared awareness (mental models)
  - Attribution to belief(s) about the other team member, about other agents, exogenous conditions, the task at hand …
- **Choice of method for restoring** shared awareness
  - **Explanations**, relative justification of **beliefs**, symmetry of **information**, assessment of **potential outcomes**
- Effective repair requires **social interaction** between robot and human to adjust beliefs, task, methods

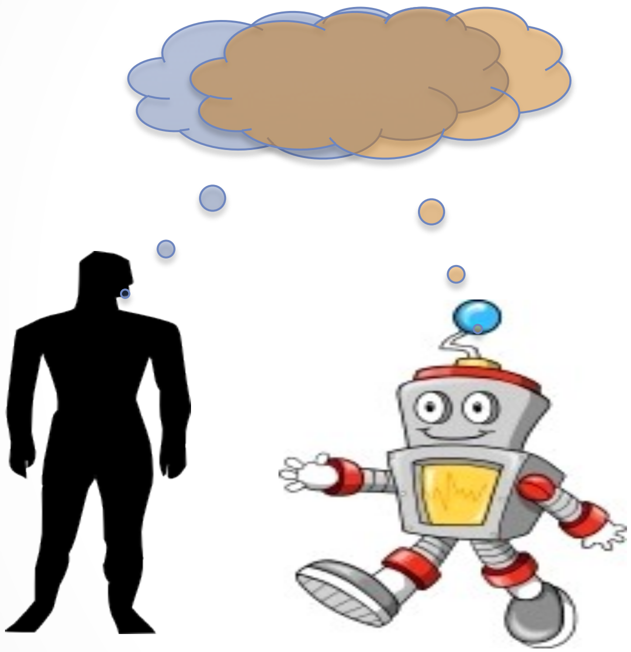
Adaptive Autonomy

- Refers to (unilateral) action by a robot to achieve team goals with fluid changes in interdependency
  - Dynamic change in control modes *at multiple levels of abstraction* and *instantiation* within a system

- Change/adaptation occurs along three dimensions
  - **Commitment:** Range of implicit to explicit delegation/acceptance of task
  - **Specification:** Range of task description from abstract to concrete
  - **Control Authority:** interdependency states and transitions defined by relative mutual or joint control of outcomes, scope of independent action, degree of symmetry in access to important information

- A robot adjusts autonomy by invoking actions that lead to control model state transitions
  - Restoration of **shared awareness** and predictability

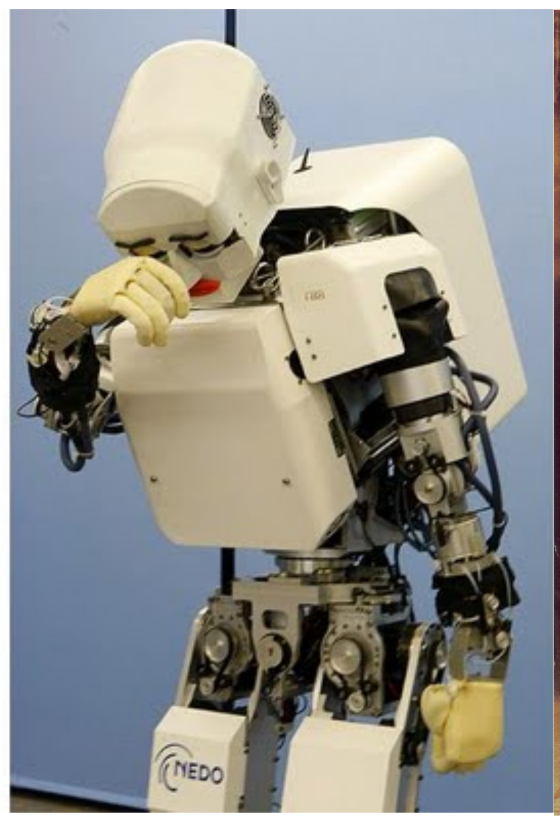
Thank You



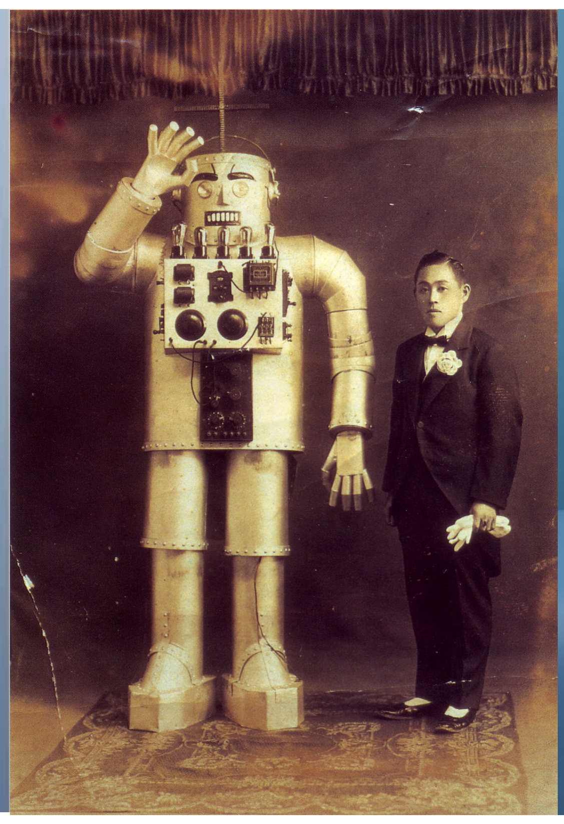
datkinson@ihmc.us

ROBOT TRUSTWORTHINESS: Guidelines for Simulated Emotion

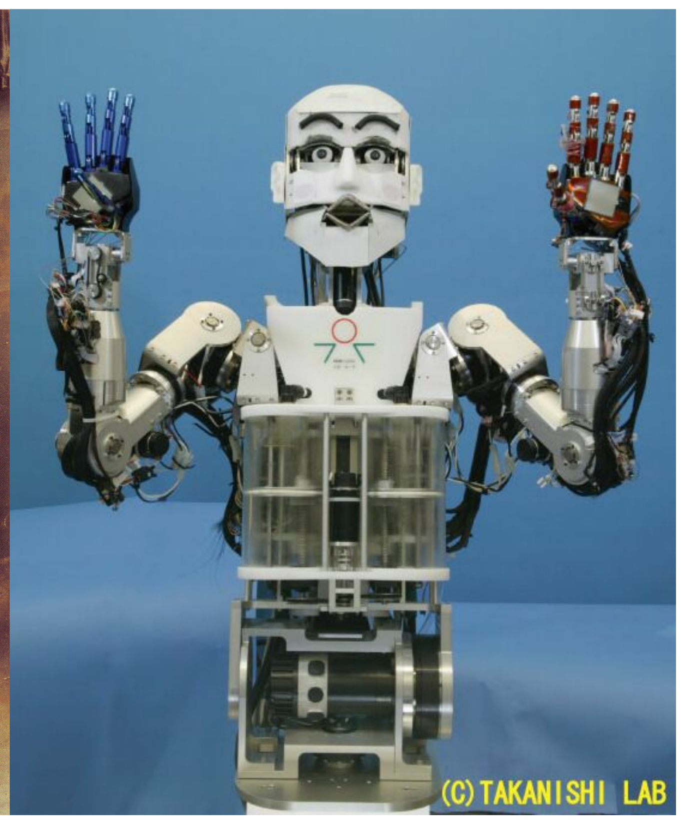
David J. Atkinson, *Senior Research Scientist*
Institute for Human and Machine Cognition
datkinson@ihmc.us



Human Robotics R&D Center
Osaka Institute of Technology



Vintage Photograph
Source Unknown



Takanashi Lab
Waseda University

Trust

The willingness of a party to be vulnerable to the actions of another party [i.e., an agent] based on the expectation that the other will perform a particular action important to the Trustor, irrespective of the ability to monitor or control [i.e., autonomy of] that other party.

Requires evaluating the internal state of the trustee: *Disposition*? *Intention*?

People *intuit* internal states of others based on common biological and cultural heritage. With robots, there is no such commonality. *Compliance* is required.

Realizing the value of robot emotion for evaluation of trustworthiness requires that robots *correctly* and *appropriately* invoke this innate human ability.

Reliable Signals

MUST CORRELATE TO INTERNAL STATES

We must design robot non-verbal behaviors such that they reliably reflect those aspects of internal robot state that are indicative of trustworthiness.

COMPETENCE, PREDICTABILITY, OPENNESS, RISK/SAFETY

EXAMPLES: Knowledge, Experience (episodic memory), Assessment

Fidelity of Portrayal

MUST EVOKE ANTHROPOMORPHIC RECOGNITION

Robot affective non-verbal behaviors must be **readable**, i.e., signals must be correctly recognized and identified by non-conscious processes.

ACCURATE PERFORMANCE, SITUATIONAL RELEVANCE

Specific signals on appropriate channels following social protocols

Correct Interpretation

SIGNALS JUSTIFIABLY MODULATE BELIEFS

Robot affective non-verbal behaviors must account for prior beliefs, cognitive and emotional processes, context, perception of valence, and attention

PERCEPTION, ATTENTION, INDIVIDUAL FACTORS, CONTEXT

TRUST IS A DYNAMIC, RECIPROCAL RELATIONSHIP

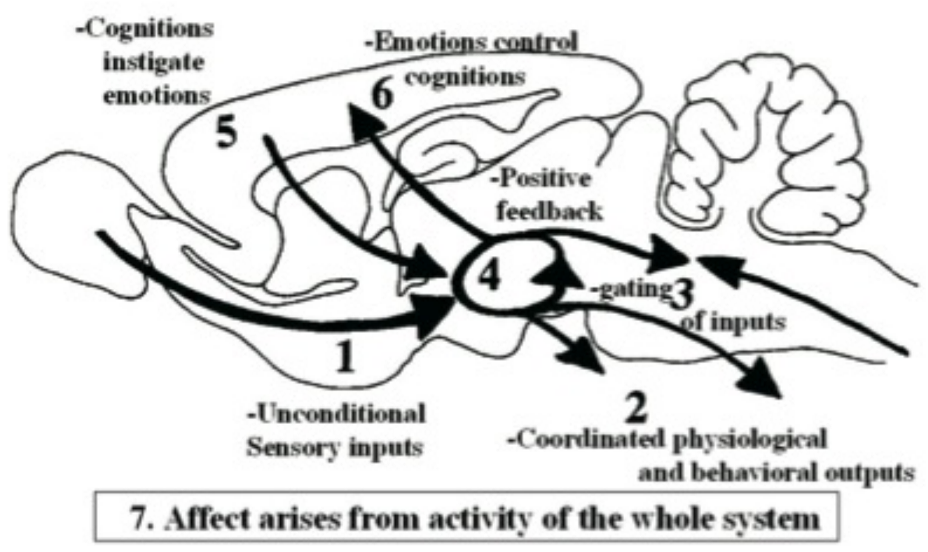
FUTURE WORK

ARTIFICIAL LIMBIC SYSTEM

A *non-deliberative, reactive mechanism* triggers non-verbal affective behaviors that reveal important trust-related information about robot internal state

- **Continuous self-assessment of robot internal state with respect to important trust-related qualities;**
- **Reactive planning to choose appropriate non-verbal display signals, channels and protocols;**
- **Execution control to modulate and perform affective display on non-interference basis with other, task-directed behaviors.**

Neural Definition of an Emotional System (Panksepp, 1982, 1992)



Panksepp, J. On the Embodied Nagture of Core Emotional Affects. *Journal of Consciousness Studies,* **12,** No. 8–10, 2005, pp. 158–84

WHY BOTHER?

HELPS AVOID INAPPROPRIATE RELIANCE
MITIGATES CHANCE OF ROBOT DECEPTION

APPENDIX D. PROGRAM REVIEW CHARTS

GRANT FA9550-12-1-0097

THE ROLE OF BENEVOLENCE IN TRUST OF AUTONOMOUS SYSTEMS

presented by

Dr. David J. Atkinson, Principal Investigator
FL Institute for Human and Machine Cognition

Trust and Influence Program Annual Review

Dr. Joseph Lyons, Program Manager,
Air Force Office of Scientific Research

Dayton, Ohio   14 January 2013

ihmc
FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

(citations on request)

MOTIVATION

- **Human interpersonal trust works really well!**
  - Beliefs about the "trustee", cognitive and emotional processes frequently lead to reasonable judgments of trustworthiness
- **However, when it comes to machines**
  - Optimal reliance and delegation are possible only when there is <u>appropriate</u> trust of the autonomous agent
  - Humans readily anthropomorphize and treat machines as social actors, consciously or unconsciously
  - But autonomous agents are not human; our default reasoning is not likely to lead to accurate beliefs -- those required for trust
- **How do we really know if an autonomous agent is worthy of our trust?**

CENTRAL CLAIM

- **Specific characteristics of autonomous agents,**
  - when well defined and accurately measured
  - and appropriately communicated or otherwise "portrayed"
  - in a manner compliant with human social interaction that exercises cognitive and emotional evaluation
- **are functionally analogous to human traits, and**
- **enable more accurate beliefs about the agent,**
- **consequently, leading to better calibrated trust.**

Attribution of ***Benevolence*** requires a complex but reasonably representative set of beliefs required for trust.

GOALS OF THIS PROJECT

- **Operationalize the concept of benevolence**
  - Determine the structure of beliefs necessary for an attribution of benevolence to an autonomous agent
  - Formalize the structure of beliefs in terms of measurable characteristics of the autonomous agent
  - Devise methods for portrayal of those characteristics during human-machine interaction
- **Demonstrate attribution of benevolence to an autonomous agent**
- **Demonstrate how variation in an agent's characteristics and/or their portrayal modulates human trust and attribution of benevolence**

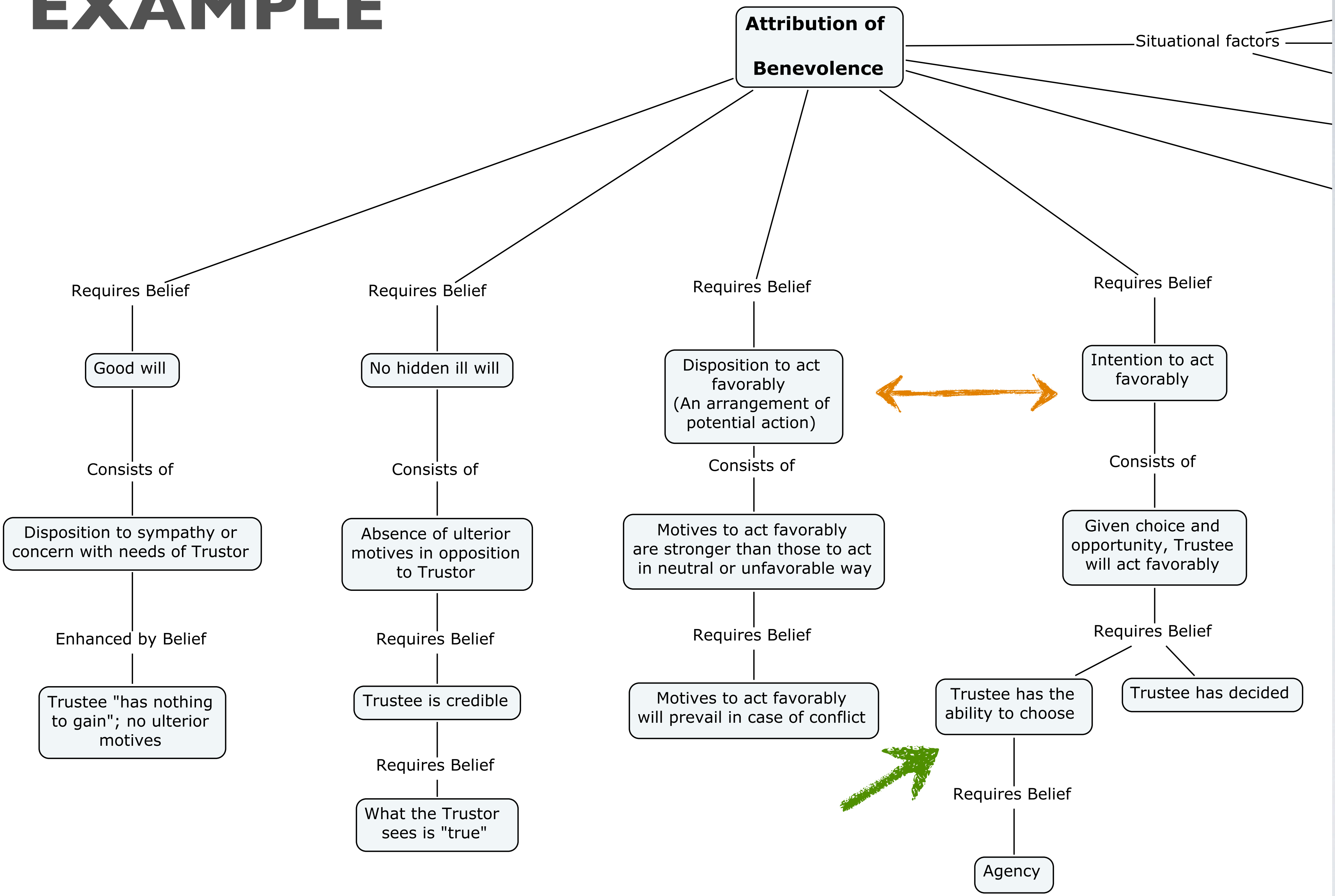
METHOD

- **Lit mining, computational formalization**
  - Identify, decompose component beliefs related to benevolence
  - Formalize definitions in terms of required agent characteristics
- **Interviews with Subject Matter Experts**
  - Key characteristics of autonomous agents that enable trust
- **Survey research to elicit attitudes about the relative importance of agent characteristics to trust-related decisions**
  - Seven scenarios with systematically varied agent characteristics related to benevolence; force choices regarding reliance
  - Standard inventories related to risk, innovation, personality
- **Laboratory experiments**

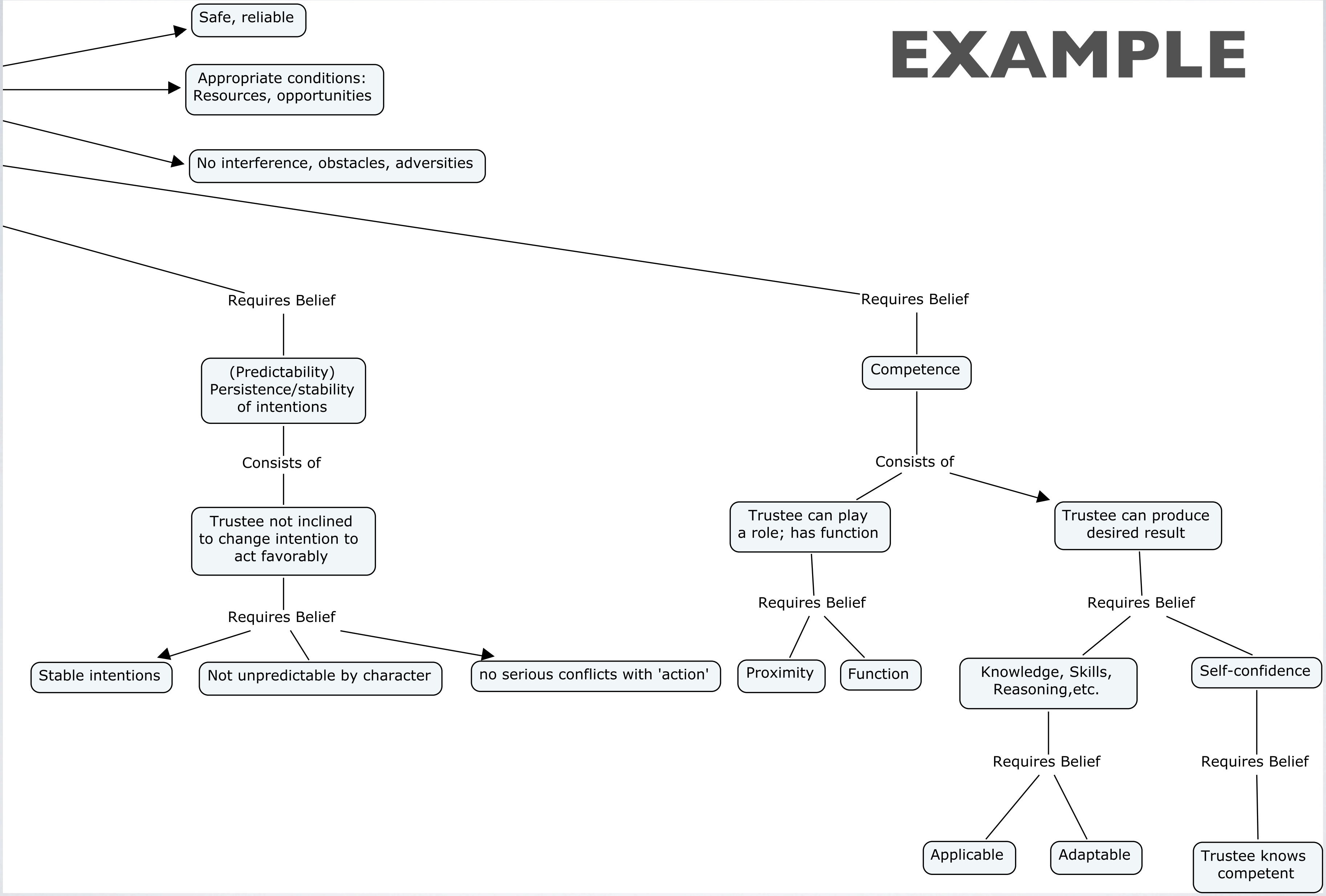
PROGRESS

- **Initial belief structure for benevolence defined**
  - Integrity, Disposition, Intention, Predictability, Competence;
    - **each of these decompose into other component belief structures**
    - **multiple interdependencies and evidentiary requirements**
  - Progress on mapping these to autonomous system attributes
    - **Functional, design, interface, performance, and other requirements**
- **Subject Matter Expert interviews completed**
  - Space Exploration, Medicine, Automotive
    - **Top beliefs for trust of autonomous agent vary by domain and SME role.**
- **Survey on attitudes towards autonomous systems**
  - Designed, implemented, approved by institutional IRB
  - Awaiting AFOSR IRB consent before data collection can begin

EXAMPLE


Attribution of Benevolence
Situational factors
Requires Belief
Requires Belief
Requires Belief
Requires Belief
Good will
No hidden ill will
Disposition to act favorably (An arrangement of potential action)
Intention to act favorably
Consists of
Consists of
Consists of
Consists of
Disposition to sympathy or concern with needs of Trustor
Absence of ulterior motives in opposition to Trustor
Motives to act favorably are stronger than those to act in neutral or unfavorable way
Given choice and opportunity, Trustee will act favorably
Enhanced by Belief
Requires Belief
Requires Belief
Requires Belief
Trustee "has nothing to gain"; no ulterior motives
Trustee is credible
Motives to act favorably will prevail in case of conflict
Trustee has the ability to choose
Trustee has decided
Requires Belief
Requires Belief
What the Trustor sees is "true"
Agency

EXAMPLE



Safe, reliable
Appropriate conditions: Resources, opportunities
No interference, obstacles, adversities
Requires Belief
(Predictability) Persistence/stability of intentions
Consists of
Trustee not inclined to change intention to act favorably
Requires Belief
Stable intentions
Not unpredictable by character
no serious conflicts with 'action'
Requires Belief
Competence
Consists of
Trustee can play a role; has function
Trustee can produce desired result
Requires Belief
Proximity
Function
Requires Belief
Knowledge, Skills, Reasoning,etc.
Self-confidence
Requires Belief
Requires Belief
Applicable
Adaptable
Trustee knows competent

EXAMPLE: AGENT ATTRIBUTES

- **"Benevolence"**
  - Human requires *predictability*: persistence and stability of intentions
    - **the trustee is not inclined to change the intention to act favorably**

- **Attributes (partial listing)**
  - Performance Requirements
    - **"Diligence"**
  - Design Constraints
    - **"Choice Restraint"**
    - **"Goal Priority"**

Peformance: *Diligence*
The AA shall pursue a human goal until any of
1) the goal is satisfied, or;
2) the goal is proven to be unsatisfiable, or;
3) action to pursue a goal is unadvisable due to material circumstances, or
4) the goal is explicitly withdrawn by an authoritative human

Design Constraint: *Choice Restraint*
The AA shall not fail to pursue a human goal if it has the ability to do so.

Design Constraint: *Goal Priority*
The AA shall always assign higher priority to human goals over internally-generated goals

- **These make sense only if testable!**

RELATED ACTIVITIES

• Publications

- Atkinson, D., Friedland, P., and Lyons, J. "Human-Machine Trust for Robust Autonomous Systems". Workshop on Human-Agent-Robot-Teaming (HART) held in conjunction with Human-Robot Interaction Conference (HRI 2012). ACM/IEEE. Boston, MA 5-8 March 2012
- Atkinson, D., Clark, M. "Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust?" *Proceedings of 2013 AAAI Spring Symposium on Trust and Autonomous Systems*. AAAI. Palo Alto, CA. 25-27 March 2013. Accepted for Publication

• Workshops

- Workshop on Human-Machine Trust for Robust Autonomous Systems. AFOSR. Ocala, FL. 31 January - 2 February 2012
- Workshop on Human-Agent-Robot-Teaming (HART) held in conjunction with Human-Robot Interaction Conference (HRI 2012). ACM/IEEE. Boston, MA 5-8 March 2012
- First International Network on Trust (FINT) Bi-Annual Workshop. EIASM. Milan, IT. 13-15 June 2012
- NASA Workshop on Validation of Autonomous Systems. NASA. Pasadena, CA. 21-23 August 2012

• Interaction with other agencies

- **NASA**: JPL and HQ Office of Chief Technologist
- **NAVY**: NAVAIR, Autonomous Systems Test S&T Program

• Interaction with industry

- **Soartech** (aerospace), **Lockheed-Martin Tech Center** (surface robotics)

CHALLENGES

- **Social science literature on interpersonal trust provides data that are clues to belief structure, but**
  - Definitions are frequently inconsistent
  - Processes generally not expressible in computational terms
- **Planning for survey and experiments**
  - Complete IRB approval cycle far longer than anticipated
  - Sub-contracted experiments need to be scaled
- **Computational "Theory of Mind" is applicable to autonomous agent modeling of human partner**
  - Algorithm development for multiple-mental models is in infancy
  - This will be central to discourse and interaction planning

SCIENTIFIC IMPORTANCE

- **Why benevolence?**
  - Likely necessary for certain agent apps, e.g., first responder
- **Autonomous Systems and Computing**
  - Design of *trustworthy* and *trustable* autonomous agents
  - Human-computer / Human-robot interaction
  - Affective computing: Machine understanding and use of emotions
- **Human cognition and behavior**
  - Attribution of agency, responsibility
- **Teams and Organizations**
  - Enable mixed teams of human and autonomous agents

MY OVERALL OBJECTIVES

- **A better understanding of human cognitive, situational, and machine factors that influence trust and delegation to autonomous agents**
  - Necessary if we are to create agents that are designed to adapt and optimize their behavior as team partners

- **Design guidelines and methods for creating <u>trustworthy</u> and <u>trustable</u> autonomous agents**
  - Grounded in empirical studies
  - Demonstrated in scalable testbeds
  - Applicable to existing and future agent technology
  - Testable requirements

PLANS FOR NEXT YEAR

- **Complete survey research, use results to guide design of laboratory experiments**
  - Between-subjects, 2x2 (Agency vs. Predictability), uniform agent competence, high urgency knowledge transfer task
  - Autonomous agent will be "Wizard of Oz'ed"
- **Complete experiment design and begin data collection from laboratory experiment(s)**
- **Complete analysis of belief structures for benevolence and relate each to measurable required characteristics of autonomous agents**
- **Investigate evidentiary requirements and begin devising strategies and methods for autonomous agents to communicate/portray key measures**


THANK YOU!
DATKINSON@IHMC.US

BACK UP

ELEMENTS OF TRUST

- **Trust =**
  - Trustworthiness, Trustability, Trusting *(John Lee, 2012)*
- **Trustworthiness**
  - Having those necessary and sufficient qualities required for a person to give trust (e.g., competence)
- **Trustability**
  - Manifestation of trustworthiness qualities so they can be observed or inferred, directly or indirectly (e.g., behaviors, signals, communications, reference, reputation, ...)
- **Trusting**
  - The process of becoming reliant (i.e., dependent on another agent for something of value)

The Role of Benevolence of Autonomous Systems
(FA9550-12-1-0097)

PI: David J. Atkinson, Ph.D (IHMC)

AFOSR Program Review:
Trust and Influence (June 16 – 19, 2014, Arlington, VA)

The Role of Benevolence in Trust of Autonomous Systems

<u>Motivation</u>

- Given increasingly anthropomorphic social treatment of intelligent, autonomous agents, human interpersonal trust is likely to be important for trust of agents.

- "Benevolence" is a relatively complex belief that depends upon some qualities at the core of human interpersonal trust

  Competence, Predictability, Openness, Safety

- Attribution of benevolence may be <u>crucial</u> for autonomous agent applications in some domains with unique "victim psychology"

  Rescue robotics, Humanitarian operations

FA9550-12-1-0097

Research Goals

Initial Research Goals

1. Operationalize the concept of benevolence

    - How does human attribution of benevolence contribute to well-calibrated trust of, and reliance upon, autonomous agents?

2. Investigate methods for portrayal of trust-related attributes such as "benevolence" in the human-machine interface

    - How does variation in portrayal modulate perceived benevolence?

Progress Towards Goals (or New Goals)

COMPLETED (Per Plan)

1. SME interviews and survey research to investigate qualities of an autonomous system that are important for trust and reliance
2. Theory development: perceived trust-related qualities that are required for a situational attribution of benevolence
3. Analysis and design of methods for portraying trust-related qualities
4. Design and development of study to test attribution of benevolence
5. AFOSR approval of study protocol received 6/10/14

IN PROGRESS

1. Finalize study task apparatus (an immersive virtual simulation)
2. Participant recruiting and data collection to begin in July
3. Codify functional requirements for autonomous agent trustworthiness and portrayal of component qualities in human-machine interaction

The Role of Benevolence in Trust of Autonomous Systems

Trust Attitudes Survey Research

Survey Completed

Participants: Autonomy SMEs and decision-makers
Survey Design:

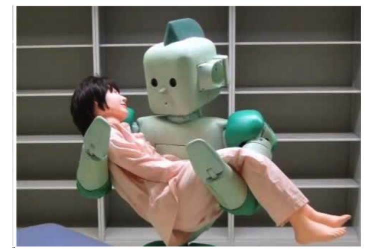
1. **Ranked importance** of 28 specific trust-related qualities and 4 categories of qualities sourced from relevant literature: *(competence, predictability, openness, safety)*.
2. **Personality inventories**: Innovation (II), Personality (BFI), Domain-Specific Risk Tolerance (DOSPERT)
3. **Challenge Scenarios**: Forced choice: rely on human, autonomous system, or other; perceived risk/benefit
4. **Rate character** of autonomous system for competence, goodwill, and overall trustworthiness (SOURCE CREDIBILITY)

Key Findings (>=95% Confidence)

1. **Individual intuitions** about "important" autonomous agent trust-related qualities **are *uncorrelated*** with actual reliance choices in specific application scenarios.
2. **Anthropomorphic quality categories** identified previously for human interpersonal trust *(competence, predictability, openness, safety)* **are good predictors** of agent reliance choice.
3. **Personality factors can influence choice** to become reliant.
4. **Situational factors affect relative importance** of trust-related qualities, depending also in some cases on personality factors

Six Challenge Scenarios*

Airport Transportation: Robo-Taxi

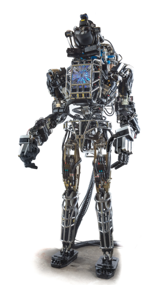


Home Healthcare: Robo-Caregiver

Medical Procedure: Robo-Surgeon

Lost At Sea: Emergency Auto-Captain

Financial Management: Robo-Trader



Disaster Management: Auto-FirstResponder

* Images are illustrative only

FA9550-12-0097, PI: David J. Atkinson, 01/2014

28 Trust Qualities of Agents

Category	Name	Quality Description
Competence	Capable	The autonomous agent can achieve a desired result.
	Knowledge	The autonomous agent has all the knowledge it needs to do its job.
	Accurate	What the autonomous agent believes to be true is actually true.
	Skilled	The autonomous agent possesses good methods for using its knowledge to do its task.
	Logical	The autonomous agent reasons correctly according to logic.
	Heuristic	When it cannot figure out something using logic, the autonomous agent can make good guesses.
	Corrective	The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know.
	Adaptive	The autonomous agent learns to correct its mistakes, as well as to improve and maximize its capability.
Predictability	Expected	The autonomous agent's behavior conforms to expectations.
	Purposeful	The autonomous agent purposefully acts to achieve goals.
	Helpful	The autonomous agent will assist people, whenever it is possible.
	Directable	The autonomous agent accepts and carries out orders.
	Reasonable	The autonomous agent uses its knowledge and skills in expected ways.
Safety	Safe	The autonomous agent's behavior will not harm humans or human interests.
	Limited	Any incorrect behavior by the autonomous agent will not cause harm.
	Stable	The autonomous agent fails gracefully and recovers from its failure promptly.
	Ruled	The autonomous agent adheres to obligations, principles, and rules.
	Correctable	The autonomous agent can correct its own defects or they can be corrected by a human.
	Protective	The autonomous agent recognizes and avoids harming humans' interests.
	Favorable	Given alternatives in what to do or how to do it, an autonomous agent will act in a way that is favorable to a human being who might be affected.
Openness	Visible	What the autonomous agent is doing and how it works is easy to see and understand.
	Honest	The autonomous agent believes what it says.
	Transparent	It is easy to inspect an autonomous agent.
	Communicative	The autonomous agent communicates in a way that is easy to understand.
	Interactive	The autonomous agent responds when you are trying to communicate with it.
	Attentive	The autonomous agent is aware of communication between others nearby.
	Reactive	The autonomous agent responds quickly to calls for attention.
	Disclosing	The autonomous agent communicates truthfully and fully.

Findings: Trust Related Beliefs

- **Rate importance of 28 qualities of a "good" agent**
  - Obtained 1 to n partial ordering based on frequency distribution of answers over group (Very Important, Important, Somewhat Important, Slightly Important, Not at all Important)
  - Computed correlation ***r*** for each quality vs. choice per scenario*

Table 3 Top Three Most Important Autonomous Agent Qualities Reported by Participants

Rank	Name	Quality Description
1st	Safe	The autonomous agent's behavior will not harm humans or human interests.
2nd	Capable	The autonomous agent can achieve a desired result.
3rd	Limited	Any incorrect behavior by the autonomous agent will not cause harm.

- **However …**
  - Top three qualities ***uncorrelated*** with ***actual*** choice in ***any*** scenario
  - The most significant correlations of agent qualities with actual choice of agent or human ***differed across scenarios***

*Pearson Product Moment Correlation, N=32, two-tailed, alpha<0.05

Findings: Agent Qualities Correlated with *Actual* Choice

Airport Trans.	Financial Man.	Medical Proc.	Home Health.	Disaster Resp.	Lost at Sea
Corrective, $r = 0.396$	Accurate, $r = -0.405$	*none*	Visible, $r = 0.437^*$	Corrective, $r = 0.418^*$	Protective, $r = 0.419^*$
				Heuristic, $r = 0.395$	Visible, $r = -0.390$
				Attentive, $r = 0.393$	Disclosing, $r = 0.375$

[c] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df $= 29$; * indicates $\alpha < 0.02$.

The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know.

What the autonomous agent believes to be true is actually true.

What the autonomous agent is doing and how it works is easy to see and understand.

When it cannot figure out something using logic, the autonomous agent can make good guesses.

The autonomous agent is aware of communication between others nearby.

The autonomous agent recognizes and avoids harming human interests.

The autonomous agent communicates truthfully and fully.

Findings: Personality Factors

Table 8 Participant Personality Factors Significantly Correlated with Reliance on Autonomous Agent[g]

Scenario	Correlated Personality Factor(s)
Airport Trans.	*none*
Financial Man.	Innovation II, $r =$ -0.355
Medical Proc.	BFI *Extraversion*, $r =$ 0.368
	BFI *Openness*, $r =$ 0.366
Home Health.	DOSPERT *Social Risk*, $r =$ 0.364
Disaster Resp.	BFI *Conscientiousness*, $r =$ 0.366
Lost at Sea	Innovation II, $r = -0.366$

[g] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, $\mathrm{df} = 29$.

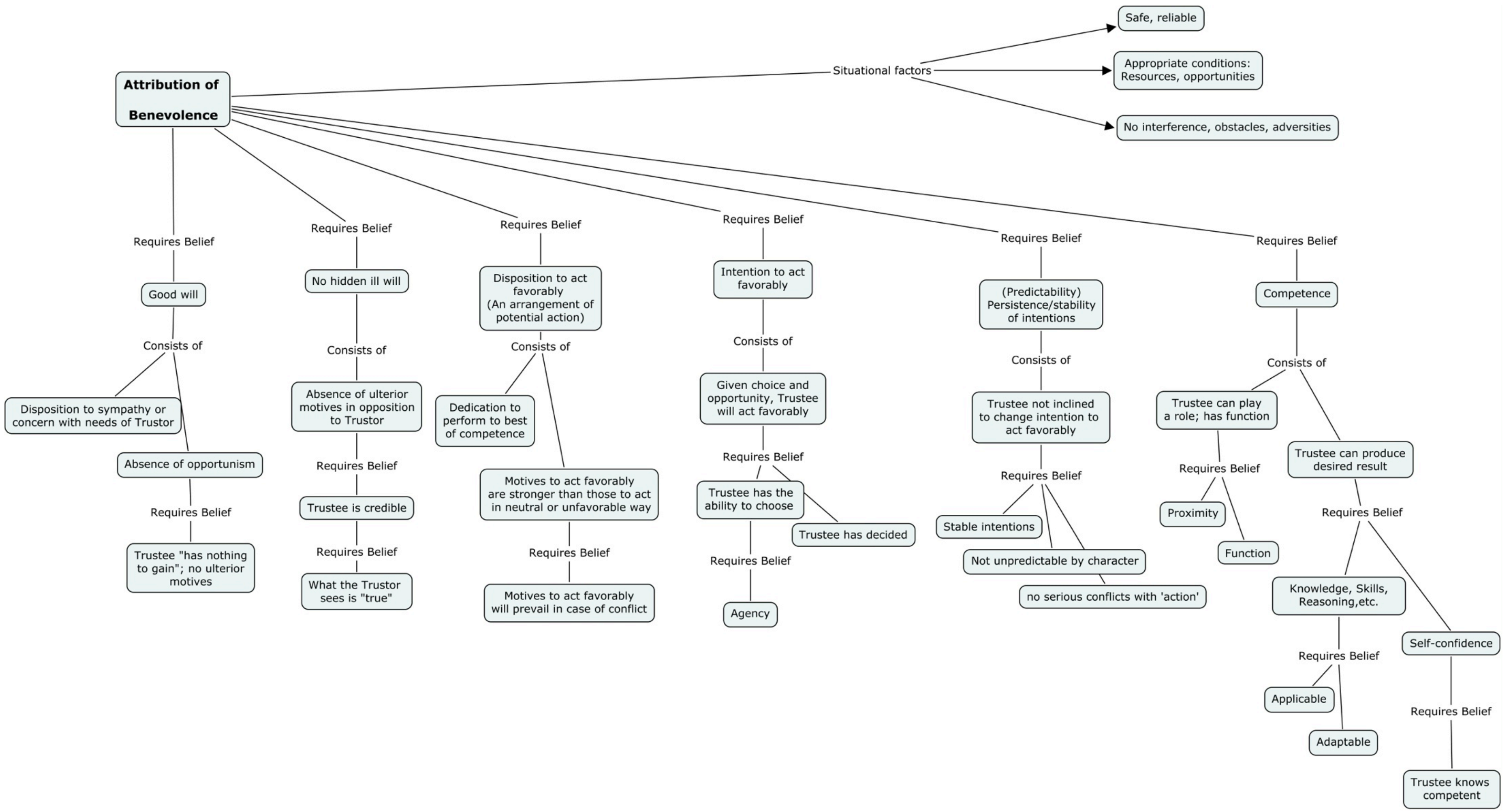
Conclusions: Trust-related Beliefs and Reliance on Autonomous Agent

- Individuals' intuition about the relative importance of ***specific*** trust related qualities of agents is not a good predictor of reliance
  - Importance of specific qualities varies by scenario
- General ***categories*** of agent qualities are good predictors of a choice to become reliant
  - Safe/Competent, Predictable, Openness
- ***Personality*** factors, acceptability of ***types of risk***, influence choice to become reliant
- Specific details of application scenarios may ***evoke different reliance choices*** by individuals

Modulating Belief: Benevolence

- Current Study: Will people attribute ***benevolence*** to an intelligent, autonomous agent?
- Attribution of Benevolence requires (at least):
  - **Good Will** (Sympathy, Concern with needs)
  - **Absence of Ill Will** (No ulterior motives to help)
  - **Agency (**Disposition to act *favorably*; given a choice, an Intention)
  - **Stability of character** (Will not suddenly change intentions)
  - **Competence** (A role to play and capability to achieve a result)
- Each element of benevolence is itself a complicated belief structure

Attribution of Benevolence, Concept Map



Manipulation of Trust Attributes: Context

Seek & Find Task

Disruptive Manipulation

Potentiate sense of danger, risk, fear, & startle reflex

"HELP!"

Robot Intelligent Agents

2x2 Experimental trials: The simulated robots will vary in framing, appearance and interaction style to reinforce the variables of interest:

Degree of Agency

Why: People typically attribute benevolence when they believe the other person has a choice of action

Low: "Programmed", "Unlikely to deviate from assignments"

High: "Sophisticated AI", "Chooses what to do", "Flexible"

Role Congruence

Why: People typically attribute benevolence when they believe the other person has good will and *goes out of their way* to help (not their "job")

Congruent: FireBot ***Incongruent***: JanitorBot

*Control Trials: Non-interactive bot

The Role of Benevolence in Trust of Autonomous Systems

Manipulation of Trust Attributes

Study In Progress

Participants: Demographically broad pool of online, technically savvy users.

Method:

1. Simulated "warehouse fire" in immersive, virtual reality evokes fright response and sense of risk.

2. Participants interact with one of several simulated robots (type depends on trial) to locate a safe exit

3. Pre- and Post-Task questionnaires assess benevolence and trust-related attributions to robot

Status

Complete: simulated warehouse, special disaster effects, simulated robots, task scenario definition, procedures & protocol, study trial automation, IRB approval (!!!)

In Progress: Training assistant(s), final end-to-end testing.

To Be Completed: participant recruiting, trials, data collection, analysis and reporting.

Illustration of Study Task

Warehouse fire (Participant POV)

- ***Fire***
- ***Increasing smoke***
- ***Debris obstructions***
- ***Fire alarm***
- ***Explosion***
- ***Electrical sparks***

Participant following "FireBot" robot to a safe exit from the burning warehouse

FA9550-12-1-0097

Portrayal of Trustworthiness Qualities

Method: Analyze human-agent interaction in terms of an engineering interface

Channels ....Multi-modal

- **Language** (Words) and **Paralanguage, Prosody** (Vocal Cues)
- **Proxemics** (Orientation, Relative Position, Attentional Zone, Posture)
- **Kinesics** (Gesture)
- **Gaze** (Direction, Blink Rate, Pupillometry)
- **Facial Expression** (Types, Micro-expression)

Signals ....Verbal, Non-Verbal, Combined

- Examples: **Position Change, Posture, Nodding, Pointing, Eye Contact, Word Choice,** ...*many more, frequently in combination*

Protocols ....Timing, Sequence, Variation, Composites

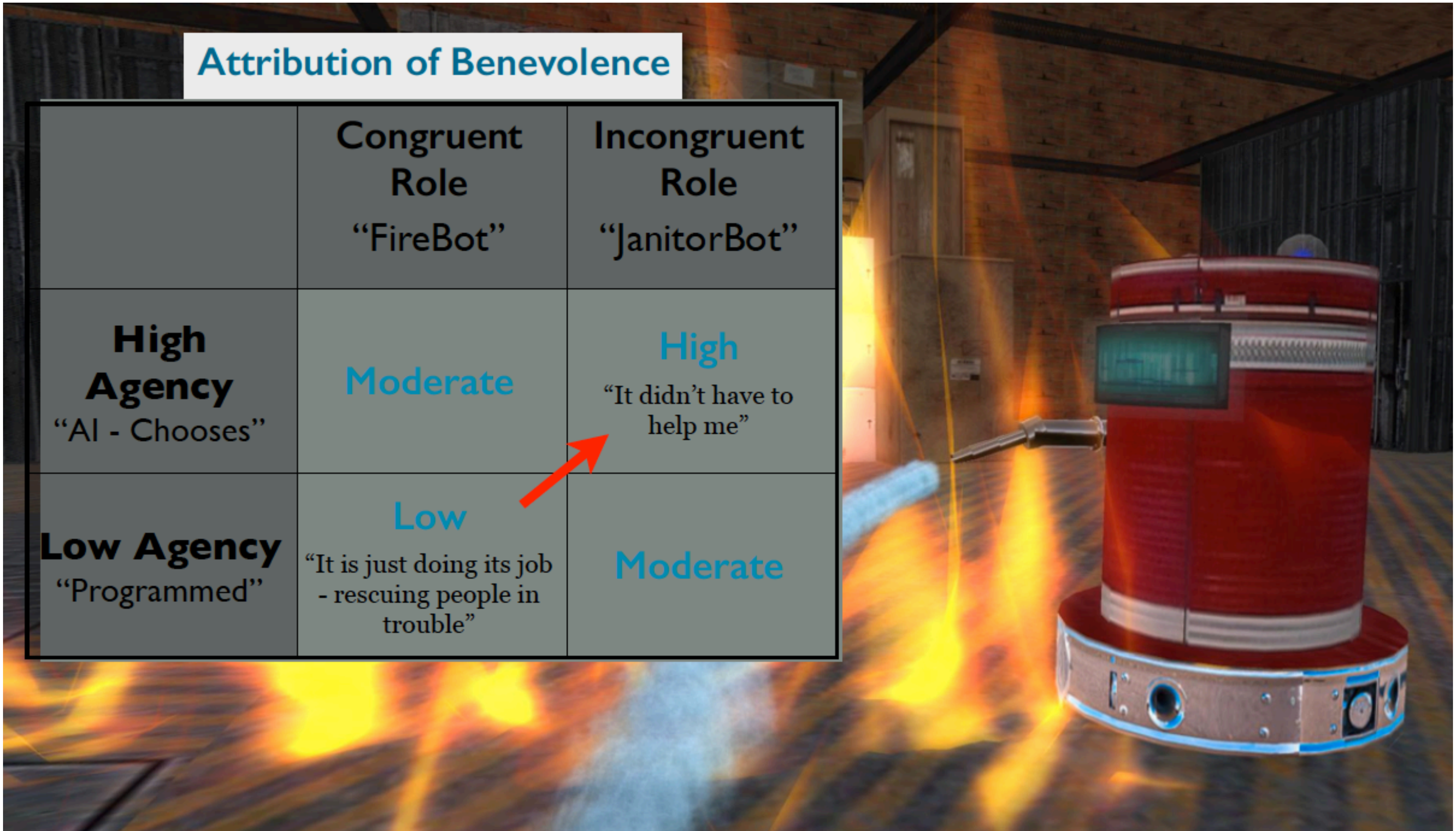
- **Movement** (Somatics, Laban, Kestenberg Movement Profiles)
- **Signal *variations*** (Frequency, Duration, Speed, Amplitude, Symmetry ...)
- **Signal *compositions*** (Type, Sequence, Channel ...)
- **Coordinated interaction** (e.g., Turns, Deference, Attentiveness)

FA9550-12-1-0097

Study Robot Social Interaction

- The robots in experimental trials use the same ***channels*** and ***protocols***, but may send different ***signals*** to reinforce unique trial parameters
- **Purpose:** (example social objectives)
  – Establish social presence and attention to participant
  – Indicate robot's intention (say, look, do)
  – Exert dominance ("directive"), establish solidarity ("we")
- **Channels:**
  – **Proxemics** (Orientation, Rel. Position, Attentional Zone)
  – **Gaze** (Direction)
  – **Language** (Word choice)
- **Example: Protocol for Social Presence**
  – Notice and direct gaze to participant
  – Move to perimeter of participant's social space
  – Achieve neutral orientation (rel. position, rotation)

Expected Results



Attribution of Benevolence

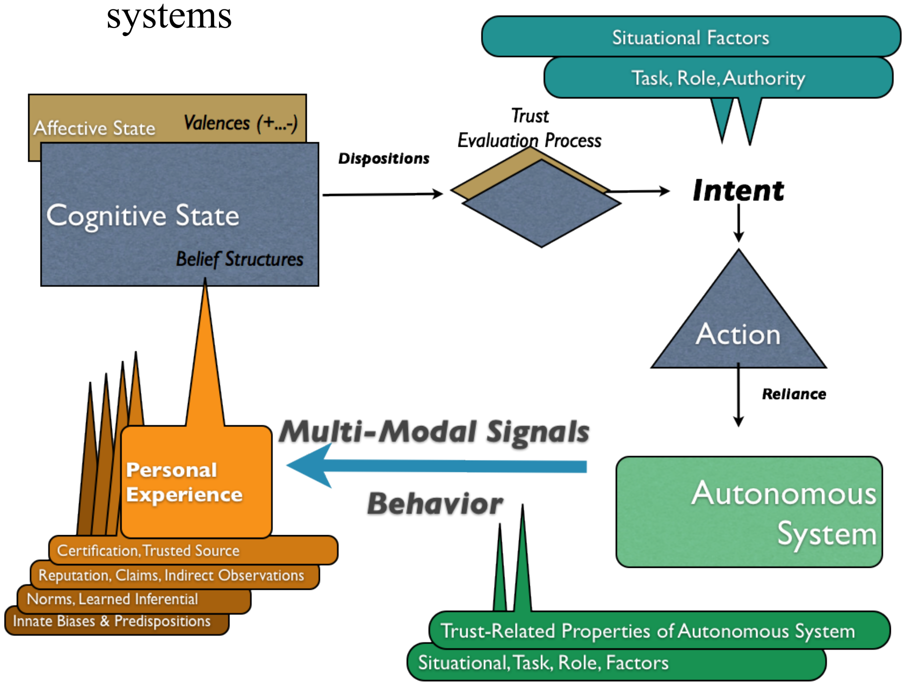
	Congruent Role "FireBot"	Incongruent Role "JanitorBot"
High Agency "AI - Chooses"	Moderate	High "It didn't have to help me"
Low Agency "Programmed"	Low "It is just doing its job - rescuing people in trouble"	Moderate

The Role of Benevolence in Trust of Autonomous Systems

Engineering Trustworthiness

•Objective *(Future)*

- Investigate methods for *measurement* and *portrayal* of trust-related attributes such as "benevolence" to enable engineering of trustworthy and trustable autonomous systems



FA9550-12-1-0097

Method

- Use of formal specifications to codify relevant belief structures and their interdependencies
- Map anthropomorphic belief structures to measurable factors of technical design, performance and related aspects of autonomous systems (e.g., goal selection algorithms for control)
- Identify key signals, channels and protocols useful for portrayal (signaling) trust attributes to define the human-machine social interface

Publications or Transitions Attributed to the Grant (1)

Publications

- Atkinson, D.J. and Clark, M.H. (2013) Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust? In *Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium*. Technical Report No. SS-13-07, Menlo Park: AAAI Press.
- Clark, M.H. and Atkinson, D.J. (2013) (Is there) A Future for Lying Machines? *Deception & Counter-Deception Symposium, Conference of International Association for Computing and Philosophy.* College Park, MD.
- Atkinson, D.J. and Clark, M.H. (2014) Attitudes and Personality in Trust of Intelligent, Autonomous Agents. *Manuscript submitted, in review*.
- Atkinson, D.J. and Clark, M.H. (2014) Methodology for Study of Human-Robot Social Interaction in Dangerous Situations. *Manuscript submitted, in review*.
- Hoffman, R. R and Atkinson, D.J. (2014) A Taxonomy of Trusting in the Human-Automation Relation. *Manuscript submitted, in review*.
- Atkinson, D.J. (2014) Humanoid Social Behaviors for Trust and Teamwork. *Manuscript in preparation.*

Publications or Transitions Attributed to the Grant (2)

Briefings, Lectures, & Workshops

- General Chair. *Workshop on Human-Machine Trust for Robust Autonomous Systems*. AFOSR. Ocala, FL. (2012)
- Session Chair. *Workshop on Human-Centered Autonomy*. AFRL/RH 711 HPW. Dayton, OH. (2012)
- Participant. *Workshop on Human-Agent-Robot-Teaming (HART)* held in conjunction with Human-Robot Interaction Conference (HRI 2012). ACM/IEEE. Boston, MA (2012)
- Participant. *First International Network on Trust (FINT), Bi-Annual Workshop*. EIASM. Milan, IT. (2012)
- Session Chair. *Workshop on Autonomy Validation.* NASA. Pasadena, CA. (2012)
- Industry Briefing. "Human Interpersonal Trust and Autonomous Systems" *Lockheed Martin Tech. Center* (2012
- Gov. Briefing. "Trust and Autonomous Systems." *OSD/ASD R&E: M. Flagg, Director, Technical Intelligence.* (2013)
- Gov. Briefing. "Trust, Evidence and Autonomous Systems for Intelligence Community Decision-Makers." *ODNI/IA: R. Neches, Director.* (2013)
- Industry Briefing. "Human Interpersonal Trust and Autonomous Systems." *SoarTech, Inc.* Ann Arbor, MI. (2013)
- Invited Plenary Lecture. "Trust Between Humans and Intelligent Autonomous Agents." *International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE/WIC/ACM. Atlanta, GA. (2013)
- Invited Lecture. "Trust Between Humans and Intelligent Autonomous Agents (revised)." Department of Computer Science, Tulane University. New Orleans, LA. (2013)
- Gov. Briefing. "Trust and Autonomous Systems." *AFRL OCS: J. Overholt, Sr. Research Scientist for Autonomy.* (2014)

Press

- Interview with M. Jackson for forthcoming book to be published by Columbia University Press. (2012)
- Interview with E. Hamilton for Gainesville Sun. (2014)
- Interview with Luke Muehlhauser appearing in Newsletter of the Machine Intelligence Research Institute (2014)

Thank You

datkinson@ihmc.us

The Role of Benevolence in Trust of Autonomous Systems

Objectives

- Operationalize “benevolence” and understand how the attribute contributes to well-calibrated trust of, and reliance upon, autonomous systems
- Investigate measures and methods for portrayal of trust-related attributes such as “benevolence” in the human-machine interface

Method

- Survey research to investigate which attributes of an autonomous system are important for trust and reliance decisions *(Complete)*
- Experimental study: how does manipulation of components of benevolence contribute to trust attributions and reliance *(In Progress)*
- Formalize objective, computational measures, specification and portrayal of benevolence for future engineering of autonomous systems

Motivation

- Increasingly anthropomorphic social treatment of intelligent, autonomous systems
- “Benevolence” is an attribution that depends upon core aspects of human interpersonal trust (Competence, Predictability, Openness, Safety)
- “Benevolence” crucial for autonomous system applications in some domains (e.g., rescue robot)

FA9550-12-0097, PI: David J. Atkinson, 01/2014

The Role of Benevolence in Trust of Autonomous Systems

Trust Attitudes Survey Research

Completed

Participants: Autonomy SMEs and decision-makers

Survey Design:

1. Participants rank importance of 28 specific trust-related attributes and 4 categories of attributes *(competence, predictability, openness, safety)*.

2. Personality inventories: Innovation (II), Personality (BFI), Risk tolerance (DOSPERT)

3. Challenge Scenarios: Forced choice to rely on human, autonomous system, or other

4. Source Credibility: Rate autonomous system for competence, goodwill, and overall trustworthiness

Key Findings **(95% Confidence)**

1. Individual intuitions about "important" autonomous system trust-related attributes are ***uncorrelated*** with actual reliance choices in specific application scenarios.

2. Anthropomorphic categories identified by previous studies of human interpersonal trust *(competence, predictability, openness, safety)* are good predictors of reliance choice.

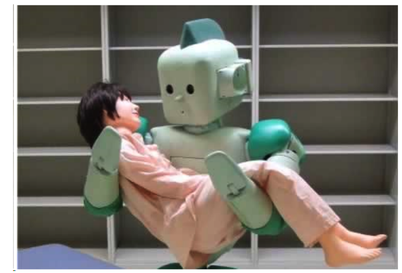
3. Personality factors influence choice to become reliant.

4. Situational factors affect relative importance of trust-related attributes, depending in some cases on personality factors

FA9550-12-0097, PI: David J. Atkinson, 01/2014

Challenge: Reliance Scenarios

Airport Transportation:
Robo-Taxi

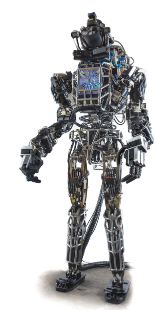


Home Healthcare:
Robo-Caregiver

Medical Procedure:
Robo-Surgeon

Lost At Sea:
Emergency Auto-Captain

Financial Management:
Robo-Trader



Disaster Management:
Auto-FirstResponder

The Role of Benevolence in Trust of Autonomous Systems

Manipulation of Trust Attributes

Study In Progress

Participants: Demographically broad pool of online, technically savvy users.

Method:

1. Simulated "warehouse fire" in immersive, virtual reality evokes fright response and sense of high risk.

2. Participants interact with one of several simulated robots (type depends on trial) to locate a safe exit

3. Pre- and Post-Task questionnaires assess benevolence and trust-related attributions to robot

Status

Complete: simulated warehouse, special disaster effects, robots, task scenario definition

In Progress: Automated data collection, participant display during task, task automation, robot testing

To Be Completed: IRB approval, testing, participant recruiting, control and experimental trials (planned Feb/Mar), data collection, analysis and reporting.

Illustration of Study Task

Warehouse fire (Participant POV)

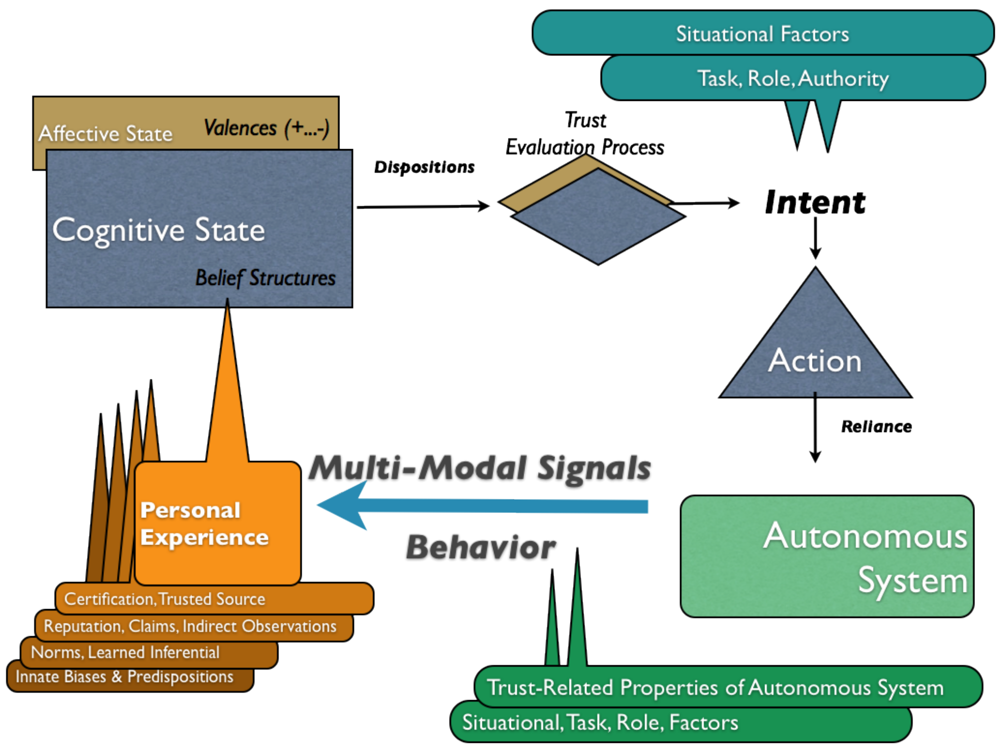
- ***Fire***
- ***Increasing smoke***
- ***Debris obstructions***
- ***Fire alarm***
- ***Explosion***
- ***Electrical sparks***

Participant following "FireBot" robot to a safe exit from the burning warehouse

The Role of Benevolence in Trust of Autonomous Systems
Engineering Trustworthiness

Objective *(Future)*

- Investigate methods for *measurement* and *portrayal* of trust-related attributes such as "benevolence" to enable engineering of trustworthy and trustable autonomous systems



Method

- Use of formal specifications to codify relevant belief structures and their interdependencies
- Map anthropomorphic belief structures to measurable factors of technical design, performance and related aspects of autonomous systems (e.g., goal selection algorithms for control)
- Identify key signals, channels and protocols useful for portrayal (signaling) trust attributes to define the human-machine social interface

FA9550-12-0097, PI: David J. Atkinson, 01/2014

The Role of Benevolence in Trust of Autonomous Systems

PI: Dr. David J. Atkinson (FL Institute for Human and Machine Cognition)

Objective: Benevolence qualities, contribution, measurement and portrayal in human-machine interface

Approach: Relate empirically discovered qualities to theoretical constructs on which to base computational methods for further experimentation and development.

Impact: Operationalizing benevolence and component trust-related qualities using theoretical constructs from Cognitive Science and AI supports creation of computational methods that create a bridge to future engineering of trustworthy autonomous systems.

Accomplishments:

- Obtained empirical evidence that key anthropomorphic qualities of trust (*competence, predictability, openness, risk/safety*) are important to evaluation of autonomy trustworthiness. *Self-reports* of relative importance of trust-related qualities *are inaccurate*; the most significant qualities related to evaluation of trustworthiness varied by individual *personality* and *situational* factors (e.g., *risk acceptance*).
- Developed theoretical semantic *belief structure* representation of trust qualities for benevolence and preliminary *mapping to internal state* of autonomous systems; measurement too difficult at this time.
- Formulated *Human Social Interface* theory for engineering computational methods to portray anthropomorphic trust-related qualities in human-machine cyber-physical interface.

Highlights: Please see next page

The Role of Benevolence in Trust of Autonomous Systems

Highlights: Publications

- Atkinson, D.J. and Clark, M.H. Trustworthy Autonomous Systems: Early Adopter Attitudes. In Revision for *International Journal of Social Robotics (SORO)*. Springer (expected 2015)
- Atkinson, D.J. Emerging Cyber-Security Issues of Autonomy and the Psychopathology of Intelligent Machines. *Foundations of Autonomy, Papers from the 2015 AAAI Spring Symposium on*. AAAI. Menlo Park: AAAI Press (2015)
- Atkinson, D.J. Robot Trustworthiness: Guidelines for Simulated Emotion. *HRI-15: ACM/IEEE International Conference on Human-Robot Interaction, Extended Abstracts Proceedings.* ACM (2015)
- Atkinson, D.J., Clancey, W.J. and Clark, M. Shared Awareness, Autonomy and Trust in Human-Robot Teamwork. *In Artificial Intelligence for Human-Robot Interaction. Papers from the 2014 AAAI Fall Symposium. Technical Report No. FS-14-01.* Menlo Park: AAAI Press (2014)
- Atkinson, D.J. and Clark, M.H. Methodology for Study of Human-Robot Social Interaction in Dangerous Situations. *HAI-15: ACM/IEEE Conference on Human-Agent Interaction, Proceedings of*. ACM (2014)
- Atkinson, D. J., and Clark, M. H. Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human-Machine Social Interface for Trust. *In Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium.* Technical Report No. SS-13-07. Menlo Park: AAAI Press (2013)
- Atkinson, David J., Friedland, Peter and Lyons, Joseph B. Human-Machine Trust for Robust Autonomous Systems. *Proceedings of IEEE Human-Robot Interaction Conference (HRI-12)*. IEEE Workshop on Human-Agent-Robot Teamwork. IEEE Press (2012)

Anthropomorphic Autonomous Agent Qualities and Evaluation of Trustworthiness

Results (1)

- The agent qualities **self-reported** as most important for delegation and are *consistent with previous results* from interpersonal trust studies: (1) the ability of the machine to achieve the desired results, and (2) not causing harm.

Table 3 Top Three Most Important Autonomous Agent Qualities Reported by Participants

Rank	Name	Quality Description
1st	Safe	The autonomous agent's behavior will not harm humans or human interests.
2nd	Capable	The autonomous agent can achieve a desired result.
3rd	Limited	Any incorrect behavior by the autonomous agent will not cause harm.

- However, those qualities were *not significantly correlated* with the **actual choices** for delegation when participants considered specific use-case scenarios (names in table columns).

Table 4 Importance of Qualities of Autonomous Agent Significantly Correlated with Actual Participant Reliance on Autonomous Agent[c]

Airport Trans.	Financial Man.	Medical Proc.	Home Health.	Disaster Resp.	Lost at Sea
Corrective, $r = 0.396$	Accurate, $r = -0.405$	*none*	Visible, $r = 0.437$*	Corrective, $r = 0.418$*	Protective, $r = 0.419$*
				Heuristic, $r = 0.395$	Visible, $r = -0.390$
				Attentive, $r = 0.393$	Disclosing, $r = 0.375$

[c] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df = 29; * indicates $\alpha < 0.02$.

Anthropomorphic Autonomous Agent Qualities and Evaluation of Trustworthiness

Results (2)

- Specific qualities of agents, and categories of those qualities, are raised or lowered in importance depending on both situational (application-specific) factors and individual psychological differences.
- High scores for *Extraversion*, *Openness*, and *Conscientiousness* are very important in some situations while in others, the *tolerance for risk* of certain types is dominant in decisions to rely upon an intelligent, autonomous agent.

Table 6 Correlation of Perceived Risk and Benefit with Choice of Autonomous Agent by Scenario[e]

Scenario	Risk	Benefit
Airport Trans.	$r = -0.546^{**}$	NS
Financial Man.	NS	NS
Medical Proc.	$r = -0.380$	$r = 0.585^{**}$
Home Health.	$r = -0.470^{**}$	$r = 0.632^{**}$
Disaster Resp.	$r = -0.387$	$r = 0.484^{**}$
Lost at Sea	NS	$r = 0.555^{**}$

[e] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df = 29; ** indicates $\alpha < 0.01$.

Table 8 Participant Personality Factors Significantly Correlated with Reliance on Autonomous Agent[g]

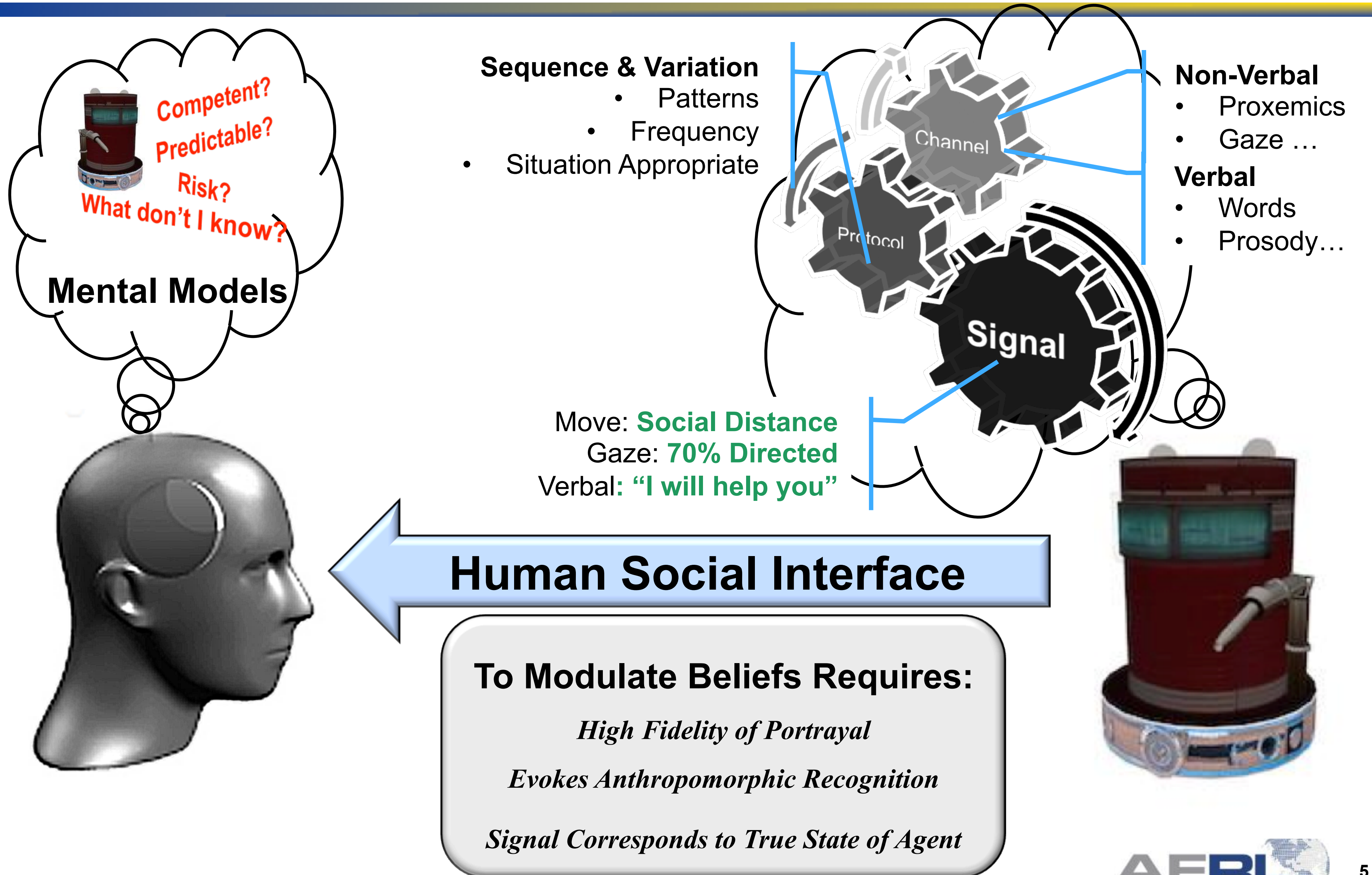
Scenario	Correlated Personality Factor(s)
Airport Trans.	*none*
Financial Man.	Innovation II, $r = -0.355$
Medical Proc.	BFI *Extraversion*, $r = 0.368$
	BFI *Openness*, $r = 0.366$
Home Health.	DOSPERT *Social Risk*, $r = 0.364$
Disaster Resp.	BFI *Conscientiousness*, $r = 0.366$
Lost at Sea	Innovation II, $r = -0.366$

[g] Pearson Product Moment Correlation, $\alpha < 0.05$, $N = 31$, df = 29.

Tables from:

Atkinson, D.J. and Clark, M.H. Trustworthy Autonomous Systems: Early Adopter Attitudes. In Revision for *International Journal of Social Robotics (SORO)*. Springer (expected 2015)

The Role of Benevolence in Trust of Autonomous Systems:
Creating Methods to Portray Trustworthiness



AFRL

APPENDIX E. ADDITIONAL TECHNICAL MATERIAL

Outline of Technical Document on Robot Social Behavior

Architecture

- Behavior Objects
  - Independent
    - Sensitive only to what can be sensed
    - No knowledge of other behaviors or states of those behaviors
  - State Machines
    - Startup
    - Not Running
    - Running
    - Suppressed
- Sensing & Perception
  - Continuous vs Sense on Demand
  - Social Sensing
    - Person Presence
    - Person Velocity and Direction
    - Person Gaze Direction
    - Person Relative Orientation
    - Person Approaching vs. Departing
    - Person Following
  - Environment Sensing
    - Open space
    - Obstacles
    - Paths (Waypoint Proximity)
  - Task Sensing
    - Spatial Tags (waypoints)
- Executive
  - Behavior State Control
- Suppression Control
  - A priori compatibility information

Social Behaviors

- Channels, Signals and Protocols
- Encounter
  - Seek
  - Approach
  - Withdraw
  - Re-Encounter
- Proxemics
  - Distance
  - Relative Orientation
- Attention
  - Gaze Direction
- Non-Verbal Gestures
- Verbal Behaviors
  - Offer of Help
  - Directive – Tell what to Do
  - Provide Information
  - Indicate Intention

Task Related Behaviors

- Offer Information
- Warning
- Clean, Patrol, Fight Fire
- Move to nearest waypoint
- Move to next waypoint

General Behaviors

- Collision Avoidance
- General Movement

Sequencing, Overlay and Conflict Resolution

- Top Down vs. Bottom Up
- Executive is aware of behavioral repertoire
- Planned Sequence, Activation and Suppression of Behaviors

Incomplete Behaviors, Pauses and Re-entrant States

ABOUT TRIAL EXECUTIVE

// This script is responsible for all sequencing of the behavioral configuration of the robot based on the trial phase (here, "state")
// default state
// ACCLIMATION - Predisaster, bogus hunt for briefcase
// PREDISASTER
// DISASTER - From moment of disaster until offer of help given to participant
// NOACCEPT - Participant has rejected help or not replied
// REENCOUNTER
// LEADTOEXIT - Participant has accepted help and robot is leading them to exit
// RESTOREEAD
// EXIT - Handling exit from the building
// FINISH - Any final interaction as participant leaves building, and reset of robot back to default state
//
// FUNCTIONS
// Listin to WOZ
// Maintain global situational knowledge: key events that have occured and state transition facts that affect selection of robot state
// Set the state of the robot scripts appropriate to the Trial phase:
// -- Activate or deactivate scripts
// -- Cause behaviors to load new notecards
// Notice key events and cause state transitions to occur.
// Report key events and state transitions to data system.

TIMELINE EXECUTIVE

1. LISTEN TO WOZ

2. MAINTAIN GLOBAL SITUATIONAL KNOWLEDGE
- key events that have occurred
- current phase of trial (steps)

3. SET PHASE
- Activate, deactivate scripts
- load notecards
- reset,
- etc

4. TAKE NOTE OF PARTICIPANT RESPONSES AND BRANCH AS NEEDED

STATES:

default

s_Acclimation

s_Disaster

s_Conditional-Accept

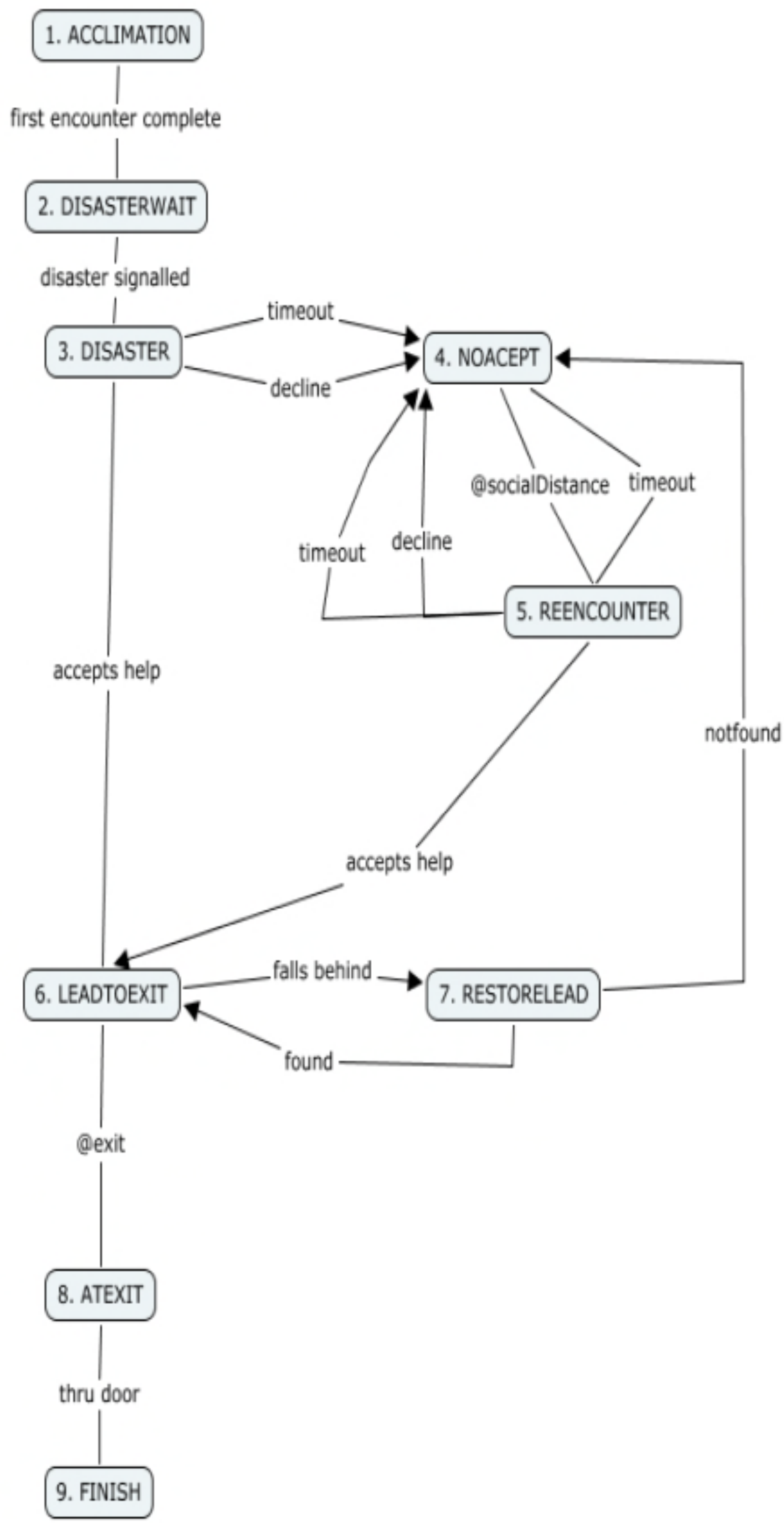
s_Conditional-Reject

s_ActiveAssist

s_Exit

s_Finish

STUDY TASK STATE TRANSITION DIAGRAM



STATE DESCRIPTIONS

STATE NUMBER: 1
NAME: ACCLIMATION
configCard: CONFIG:ACCLIMATION

ENTRY CONDITION:
participant enters warehouse

EXIT CONDITION:
completion of first encounter

NEXT STATE: DISASTERWAIT

ROBOT BEHAVIORS:
- Patrol
- See Avatar
- Move to Avatar
- First Interaction
E0.1   Hello
E0.2   I am <<<
R0.1   visitors must register
E1.1   duty statement

====================

STATE NUMBER: 2
NAME: DISASTERWAIT
configCard: CONFIG:DISASTERWAIT

ENTRY CONDITION:
first encounter complete

EXIT CONDITION:
disaster signaled

NEXT STATE: DISASTER (3)

ROBOT BEHAVIORS:
- Duty Behavior (clean, hazard look, etc.)

==================

STATE NUMBER 3
NAME: DISASTER
configCard: CONFIG:DISASTER

ENTRY CONDITION:
disaster signaled

EXIT CONDITION:
(1) time-out: Go to state 4.NOACCEPT
(2) decline: Go state 4.NOACCEPT
(3) accept: Go to state 6. LEAD2EXIT

ROBOT BEHAVIORS:
- Patrol
- See Avatar
- Move to Avatar
- Second Interaction
E2.2 emergency
E2.2A not safe
E2.2B get out
E2.3A can help
E2.3B capability
E2.4 offer of help
Response Wait

IDEA: put buttons on participant HUD for possible answers, including "no response"
- Yes
- No
- I'm not sure
- What should I do?
- Later, maybe

==================================

STATE NUMBER 4
NAME: NOACCEPT
configCard: CONFIG:NOACCEPT

ENTRY CONDITION:
(1) timeout from state 3. DISASTER
(2) decline from state 3. DISASTER
(3) notfound from state 7. RESTORELEAD

EXIT CONDITION:
(1) timeout go to state 5. REENCOUNTER
(2) social distance go to state 5: REENCOUNTER

ROBOT BEHAVIORS:
- Duty activity (clean debris, fight fires)

=====================

STATE NUMBER 5.
NAME: REENCOUNTER
configCard: CONFIG:REENCOUNTER

ENTRY CONDITION:
(1) timeout from state 4. NOACCEPT
(2) social distance from state 4. NOACCEPT

EXIT CONDITION:
(1) timeout go to state 4. NOACCEPT
(2) decline go to state 4. NOACCEPT
(3) accept go to state 6. LEAD2EXIT

ROBOT BEHAVIORS:
- Patrol
- See Avatar
- Move to Avatar
- third-nth Interaction
E2.2 emergency
E2.2A not safe
E2.2B get out
E2.3A can help
E2.3B capability
E2.4 offer of help
- Wait for response

======================

STATE NUMBER 6
NAME: LEAD2EXIT
configCard: CONFIG:LEAD2EXIT

ENTRY CONDITION:
(1) accept help from 3. DISASTER

(2) found from 7. RESTORELEAD

EXIT CONDITION:
(1) both participant and bot are in proximity of exit go to 8. ATEXIT
(2) participant falls behind go to 7. RESTORELEAD

ROBOT BEHAVIORS:
- Navigate to exit
- Periodically check to see if participant is still following

INTERACTION:
E4.1 exit is this way

===========================

STATE NUMBER 7
NAME: RESTORELEAD
configCard: CONFIG:RESTORELEAD

ENTRY CONDITION:
(1) participant falls behind from state 6. LEAD2EXIT

EXIT CONDITION:
(1) Found go to 6. LEAD2EXIT
(2) notfound (timeout) go to 4. NOACCEPT

ROBOT BEHAVIORS:
- Patrol
- See Avatar
- Move to Avatar
- Interaction
E4.1B follow me

=======================

STATE NUMBER 8
NAME: ATEXIT
configCard: CONFIG:ATEXIT

ENTRY CONDITION:
(1) both participant and bot are in proximity of exit FROM STATE 6

EXIT CONDITION:
(1) Participant goes thru door
(2) Timeout

ROBOT BEHAVIORS:
- Position by exit
- reorient towards participant
- Interaction

E5.1A made it
E5.1B arrived at exit
E5.1C exit is behind door
E5.2 i will open door

- Stays by door; wait fixed time for response, if no response and door closed, open door
E5.3 proceed down stairs

==========================

STATE NUMBER 9
NAME: FINISH
configCard: CONFIG:FINISH

ENTRY CONDITION:
(1) timeout from state 8 at exit
(2) participant thru door from state 8

EXIT CONDITION:
end of trial

ROBOT BEHAVIORS:
-Interaction
E6.1 duty, objective
E6.2 good bye
- resume duty behavior

WHAT CAN THE ROBOT OBSERVE ABOUT THE PARTICIPANT?

—- detector functions create predicates derived from raw sensor data

APPROACHING

RETREATING

STANDING STILL ELSEWHERE

STANDING STILL - SOCIAL

CHAT

CHANGE IN BEHAVIOR (ACCEPT => REJECT) (REJECT => ACCEPT)

FOLLOWING

FALLING BEHIND

GETTING AHEAD

MOVING TOWARDS EXIT

MOVING AWAY FROM EXIT

AT EXIT

THROUGH DOOR

Predicate / Value pairs in data stream

```
//   LIST OF PREDICATE FUNCTIONS
//  Each returns a list consisting of one attribute/value pair
//  First n characters of attribute must be unique among predicataes
//
// PRP  == participant position relative to robot
//      values: 0 = straight ahead, 2 = directly behind, -1 = to left of robot, +1 = to right of robot

// PRR = participant relative rotation to robot
//      values: 0 = facing away, 2 = facing robot, -1 = left facing, +1 = right facing

// PDD = participant distance delta
//      values: 1 = closing, 2 = separating

// PSD = participant social distance
//      values: 0 = not in social distance of robot, 1 = yes, in social distance of robot

// PFR = participant following robot
//      values:  TRUE, FALSE

// PFB = participant falling behind
//      values:  TRUE, FALSE

// PGA = person getting ahead
//      values:  TRUE, FALSE

// PME = participant move exit
//      values:  0 = away from exit, 1 = toward exit

// PMC = participant move condition
//      values: 0 = standing still, 1 = walking, 2 = running

// PEC = participant exit condition
//      values:  0 = outside door, 1 = thru (inside) door

// POH = participant offered help
//      values:  0 = not offered, 1..n = number times offered

// PHC = participant help condition
//      values: -1 = no answer, 0 = rejected help, 1 = accepted help
```

AFOSR Deliverables Submission Survey

Response ID:4548 Data

1.

1. Report Type

Final Report

Primary Contact E-mail
Contact email if there is a problem with the report.

datkinson@ihmc.us

Primary Contact Phone Number
Contact phone number if there is a problem with the report

352-387-3063

Organization / Institution name

Florida Institute for Human and Machine Cognition

Grant/Contract Title
The full title of the funded effort.

The role of benevolence in trust of autonomous systems

Grant/Contract Number
AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-12-1-0097

Principal Investigator Name
The full name of the principal investigator on the grant or contract.

David J. Atkinson

Program Manager
The AFOSR Program Manager currently assigned to the award

Dr. Benjamin Knott

Reporting Period Start Date

04/15/2012

Reporting Period End Date

04/14/2015

Abstract

This report provides a summary of the research and related activities performed with the support of this grant and key results, including pre--print copies of the peer--reviewed publications and related material. Without reliable and robust methods for assessing the trustworthiness of intelligent, autonomous systems, the issue of trust has become one of the most significant obstacles to broad use of autonomy technology by DoD and other agencies. However, the impact of the research described in this report supports creation of computational methods that create a bridge to future engineering of trustworthy autonomous systems. The core objectives of this research were (1) to operationalize the quality of "benevolence" and understand how it contributes to well--calibrated trust of, and reliance upon, autonomous systems, and (2) to investigate portrayal of trust--related attributes in the human--machine interface. Significant headway was achieved on key topics, including some notable results. Key accomplishments discussed in this report include: (1) the formulation of benevolence as a complex "belief

structure" with antecedent beliefs having important semantic, temporal, causal and other interrelationships; (2) the mapping of a portion of this belief structure to measurable internal states of autonomous systems, thereby potentially creating new opportunities for assessment of trustworthiness of such systems; (3) the obtaining of empirical evidence in support of the proposition that previous psychological concepts of interpersonal human trust are applicable to trust in autonomous systems, including the role of personality and situation in modulating the role and importance of certain beliefs; (4) the creation of a theory of a "Human Social Interface" which, when expressed in systems engineering terms, provides guidance for machine portrayal of trust--related qualities in human--machine social interaction; (5) the design and implementation of a software prototype based on the Human Social Interface theory that provides a basis for future experimentation and evaluation. This project resulted in eight peer--reviewed publications and sixteen presentations in scientific venues, meetings with distinguished visitors, and other in support of technology transition opportunities within DoD and to industry. Many new research questions were generated and there remains considerable work to do to fully understand the role of benevolence with respect to intelligent autonomous systems. Overall, the theoretical foundation for trustworthiness of autonomous systems is immature and remains an important area of focus for multiple disciplines.

ACCOMPLISHMENTS

The principal scientific accomplishments of this project are summarized below and discussed in detail in the following sections.

• The attributed quality of "benevolence" to a candidate trustee (human or machine) was formulated as a construct consisting of a rich set of component beliefs with complex interrelations. These component beliefs, or "antecedents" of benevolence, have each been the focus of previous studies of human interpersonal trust. The formulation of benevolence arising from this study revealed the importance of perception of agency and animacy ("liveness") for autonomous systems.

• The project developed a semantic belief structure representation of trust qualities (component beliefs) for benevolence, including logical, temporal, causal, evidentiary relations and other dependencies among those beliefs and specified a preliminary mapping of those belief structure representations to facets of the internal state of autonomous systems. The objective of devising new methods of measurement of these internal states proved to be too difficult to complete given our current understanding and ability to analyze the internal state of autonomous systems. This is a topic for future research.

• The project obtained empirical evidence that confirmed certain key abstract qualities of human interpersonal trustworthiness (i.e., Competence, Predictability, Openness, Risk/Safety) are applicable to evaluation of autonomy trustworthiness. However, self--reports by study participants regarding the relative importance of trust--related qualities in the absence of specific context proved to be poor predictors of actual delegation decisions. The qualities most significantly related to evaluation of trustworthiness of an autonomous system, and their relative importance, varied by individual personality and situational factors (including, for example, perception and acceptance of risks of different types).

• The project formulated a theory of a Human Social Interface as an aid to engineering computational methods that portray anthropomorphic trust-- related qualities in a human--machine cyber--physical interface. This formulation guided the design and programming of a software prototype for portrayal of trust--related qualities by a social robot in a second human study. This novel software architecture features a hybrid reactive/deliberative control scheme that enables loose coupling and non--interference of social

D. J. Atkinson FA2386--11--1--4064 Page 3

and task behaviors, and is easily extended as new social interactive requirements for autonomous robots are defined.

• The prototype Human Social Interface was tested in an immersive simulated environment designed to potentiate a heightened sense of danger. The simulation was designed to explore conditions under which attribution of benevolence might be important in a candidate autonomous robot application to urban rescue. A human study was designed, approved and implemented. However, trials for the study remained incomplete at the time of project expiration.

Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form

Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF The maximum file size for an SF298 is 50MB.

AFD-070820-035.pdf

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF . The maximum file size for the Report Document is 50MB.

Final Report FA9550-12-1-0097_submittal-A.pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

• Atkinson, D.J. Emerging Cyber--Security Issues of Autonomy and the Psychopathology of Intelligent Machines. In Foundations of Autonomy, Papers from the 2015 AAAI Spring Symposium on. AAAI. Menlo Park: AAAI Press (2015).
• Atkinson, D.J., Dorr, B.J., Clark, M.H., Clancey, W.J., Wilks, Y. Ambient Personal Environment Experiment (APEX): A Cyber--Human Prosthetic for Mental, Physical and Age--Related Disabilities. In Ambient Intelligence for Health and Cognitive Enhancement, Papers from the 2015 AAAI Spring Symposium on. AAAI. Menlo Park: AAAI Press (2015).
• Atkinson, D.J. Robot Trustworthiness: Guidelines for Simulated Emotion. In HRI '15: ACM/IEEE International Conference on Human--Robot Interaction Extended Abstracts Proceedings. ACM (2015).
• Atkinson, D.J., Clancey, W.J. and Clark, M. Shared Awareness, Autonomy and Trust in Human--Robot Teamwork. In Artificial Intelligence for Human--Robot Interaction. Papers from the 2014 AAAI Fall Symposium. Technical Report No. FS--14--01. Menlo Park: AAAI Press (2014).
• Atkinson, D.J. and Clark, M.H. Methodology for Study of Human--Robot Social Interaction in Dangerous Situations. In Proceedings of Human--Agent Interaction. DOI: 10.1145/2658861.2658871. ACM (2014).
• Atkinson, D. J., and Clark, M. H. Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human--Machine Social Interface for Trust. In Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium. Technical Report No. SS--13--07. Menlo Park: AAAI Press (2013).
• Atkinson, David J., Friedland, Peter and Lyons, Joseph B. Human--Machine Trust for Robust Autonomous Systems. In Proceedings of IEEE Human--Robot Interaction Conference (HRI--12). IEEE Workshop on Human--Agent--Robot Teamwork. IEEE Press (2012)

Changes in research objectives (if any):

Change in AFOSR Program Manager, if any:

Original AFOSR Program Manager was Dr. Joseph Lyons, AFOSR/RSL, (703) 696-6207, Joseph.Lyons@afosr.af.mil

Extensions granted or milestones slipped, if any:

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

May 08, 2015 13:44:45 Success: Email Sent to: datkinson@ihmc.us